Topics in experimental and tournament design

Topics in experimental and tournament design

A dissertation presented
by

Jonathan Hennessy

to
The Department of Statistics
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of

Statistics

Harvard University
Cambridge, Massachusetts
August 2014

Dissertation Advisors: Professors Tirthankar Dasgupta and Mark Glickman Jonathan Hennessy

Topics in experimental and tournament design

Abstract

We examine three topics related to experimental design in this dissertation. Two are related to the analysis of experimental data and the other focuses on the design of paired comparison experiments, in this case knockout tournaments. The two analysis topics are motivated by how to estimate and test causal effects when the assignment mechanism fails to create balanced treatment groups. In Chapter 2, we apply conditional randomization tests to experiments where, through random chance, the treatment groups differ in their covariate distributions. In Chapter 4, we apply principal stratification to factorial experiments where the subjects fail to comply with their assigned treatment. The sources of imbalance differ, but, in both cases, ignoring the imbalance can lead to incorrect conclusions.

In Chapter 3, we consider designing knockout tournaments to maximize different objectives given a prior distribution on the strengths of the players. These objectives include maximizing the probability the best player wins the tournament. Our emphasis on balance in the other two chapters comes from a desire to create a fair comparison between treatments. However, in this case, the design uses the prior information to intentionally bias the tournament in favor of the better players.

Contents

Acknowledgments

This dissertation would not have been possible without Tirthankar Dasgupta, Mark Glickman, and Luke Miratrix. I am especially grateful that Tirthankar, Luke, and Cassandra Pattanayak included me in the product marketing project that led to Chapter 2 and to have worked with Mark on the tournament design project in Chapter 3. I would also like to acknowledge Don Rubin, who inspired and guided our work in Chapter 4, and whose wisdom and ideas have heavily influenced this dissertation.

I would also like to thank my friends in the department and my family. My parents have provided so much love and support and I am blessed to have shared this experience with my sister. Finally, I would like to thank Valeria, my partner in this and in everything.

To my mom, dad, and sister

Chapter 1

Introduction

We examine three topics related to experimental design in this dissertation. Two are related to the analysis of experimental data and the other focuses on the design of paired comparison experiments, in this case knockout tournaments. The two analysis topics are motivated by how to estimate and test causal effects when the assignment mechanism fails to create balanced treatment groups. In Chapter 2, we apply conditional randomization tests to experiments where, through random chance, the treatment groups differ in their covariate distributions. In Chapter 4, we apply principal stratification to factorial experiments where the subjects fail to comply with their assigned treatment. The sources of imbalance differ, but, in both cases, ignoring the imbalance can lead to incorrect conclusions. We next review the following three chapters in more detail.

In Chapter 2, we first review the history of the conditional randomization test, which was introduced by Cox (1982) in the context of sequential designs. Then, following Rosenbaum (2002) and Zheng and Zelen (2008), we use the test as a simple

but flexible form of covariate adjustment by conditioning on the observed covariate balance. We introduce new notation to describe covariate balance and prove that the conditional test has the correct significance level. We consider a variety of sampling approaches to sample alternative treatment assignments and explore connections to rerandomization (Morgan and Rubin, 2012). Through simulation, we evaluate the properties of the conditional test and, finally, apply it to data from a product marketing experiment.

In Chapter 3, we present a methodology for finding optimal knockout tournament designs when the strengths of the players, $\boldsymbol{\theta}$, are uncertain. Following Glickman (2008), we model player strengths using a multivariate normal distribution, $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We consider 4, 8, and 16 player knockout tournaments and find the tournament brackets that maximize a variety of utility functions, for instance, the probability the best player wins the tournament and the probability the two best players meet in the final. Our emphasis on balance in the other chapters comes from a desire to create a fair comparison between treatments. However, in this case, the design uses the prior information to intentionally bias the tournament in favor of the better players. We apply Bayesian optimal design approaches, including Monte Carlo integration and simulated annealing, to identify the optimal tournament bracket. We also compare the optimal brackets to other knockout tournament designs, including brackets following the standard seeding.

In Chapter 4, we consider 2^2 factorial experiments when one of the factors is subject to all-or-nothing compliance. We use the principal stratification framework to define the finite-population and super-population factorial effects for the different

strata and use a Bayesian model to estimate the effects. We also carry out a simulation study to compare different approximations of the posterior distributions and discuss how different assumptions can be relaxed and the computational consequences.

Chapter 2

Conditional randomization tests in randomized experiments

2.1 Introduction

Randomized experiments are the "gold standard" for assessing causal effects. Randomization removes experimental bias, allows for unbiased estimation of average causal effects, and gives a "reasoned basis for inference" (Fisher, 1935). Covariates are often collected in randomized experiments and while randomization ensures that these covariates will be balanced on average, chance imbalances do occur. To quote Senn (1989),

> A frequent source of anxiety for clinical researchers is the process of randomization, and a commonly expressed worry, despite the care taken in randomization, is that the treatment groups will differ with respect to some important prognostic covariate whose influence it has proved impossible to control by design alone.

For the imbalance to be an issue, the covariate needs to be prognostic (i.e. related to the outcome) but the covariate imbalance does not need to be statistically significant in order to affect the results of the trial (Altman, 1985). Also, Senn (1989) argued that in hypothesis testing "covariate imbalance is of as much concern in large studies as in small ones" because "it is not the absolute imbalance which is important but the standardized imbalance and this is independent of sample size."

Restricted randomization and blocking are well-established strategies to ensure balance on key covariates. More recently, Morgan and Rubin (2012) introduced rerandomization as a way to ensure balance on many covariates. However, restricted randomization, blocking, and rerandomization are not always feasible. In the product marketing example that motivated this work, the covariate information was not collected until after the units were assigned to treatment levels. The experiment involved roughly 2000 experimental subjects and each subject randomly received by mail one of eleven versions of a particular product. Each subject used the product and returned a survey regarding the product's performance. The outcome of interest was an ordinal variable with three levels, 1, 2, and 3, and the goal was to identify which product version the subjects preferred. The survey also collected covariate information, including income and ethnicity and the experimenters were concerned about the effect of covariate imbalance on their conclusions.

While several methods exist to analyze ordinal data, including the proportional odds model, randomization tests are a natural choice because they require no assumptions about the distribution of the outcome. Randomization tests are unique in statistics in that inference is completely derived from the physical act of random-

ization. However, historically, randomization tests have not been used to adjust for covariate imbalance. To quote Rubin (1980),

> More complicated questions, such as those arising from the need to adjust for covariates brought to attention after the conduct of the experiment ... require statistical tools more flexible than FRTED (Fisher randomization test for experimental data).

In what follows, we explore conditioning as a way to adjust randomization tests for covariate imbalance. Conditional randomization tests have traditionally been used in the sequential design literature and only occassionally for covariate adjustment (Rosenbaum, 2002; Zheng and Zelen, 2008). In Section 2.2, we review the notation and basic mechanics of randomization tests. In Section 2.3, we introduce conditional randomization tests and prove that the test has the correct significance level. In Section 2.4, we apply the conditional randomization test to experiments with covariates. In Section 2.5, we evaluate the properties of the conditional randomization test via simulation and, in Section 2.6, we apply the test to the product marketing example. In Section 2.7, we summarize our findings and lay out steps for future work.

2.2 Randomization tests

As mentioned earlier, Fisher (1935) introduced the randomization tests for randomized experiments and the tests have played a fundamental role in the theory and practice of statistics. The early theory was developed by Pitman (1938) and Kempthorne (1952). In fact, Kempthorne (1952) showed that many statistical procedures can be viewed as approximations of randomization tests. To quote Bradley (1968),

> Eminent statisticians have stated that the randomization test is the truly correct one and that the corresponding parametric test is valid only to the extent that it results in the same statistical decision.

We introduce the notation and briefly review the mechanics of randomization tests. We then prove that the test has significance level α.

2.2.1 Notation

Let N be the number of experimental units and let K be the number of treatment levels. Each experimental unit is assigned to one level of the treatment and let $\mathbf{W} = (W_1, \ldots, W_N)$ be the treatment assignment vector. We initially consider cases where $K = 2$ and we refer to the two treatment levels as treatment and control. This assumption though is not critical to our discussion. If $W_i = 1$, unit i is assigned to treatment and if $W_i = 0$, unit i is assigned to control. Let $N_T = \sum_{i=1}^{N} W_i$ be the number of treated units and $N_C = N - N_T$ the number of control units. Let Ω be the set of all possible treatment assignments. In the case of two treatment levels, $|\Omega| = 2^N$. In the case of K treatment levels, $|\Omega| = K^N$. In selecting the actual treatment assignment, we typically consider a restricted set of treatment assignments, which we deem acceptable. For instance, a completely randomized design restricts the set to treatment assignments where N_T is fixed. We call the set of *acceptable treatment assignments* $\mathscr{S}$, where $\mathscr{S} \subseteq \Omega$. Let $\mathrm{p}(\mathbf{W})$ be the probability treatment assignment $\mathbf{W}$ is selected. We implicitly condition on $\mathbf{W} \in \mathscr{S}$ such that $\mathrm{p}(\mathbf{W}) = \mathrm{p}(\mathbf{W} \mid \mathbf{W} \in \mathscr{S})$. Traditionally, all treatment assignments in $\mathscr{S}$ are equally likely. For instance, in the completely randomized design, $\mathrm{p}(\mathbf{W}) = \binom{N}{N_T}^{-1}$. This assumption, though, is also not critical.

2.2.2 Randomization test mechanics

The randomization test tests the Fisher sharp null hypothesis of no treatment effect, which can be written as $H_0 : Y_i(1) = Y_i(0)$ for $i = 1, \ldots, N$. The test requires the experimenter to make two choices. The first is the choice of the test statistic, $t(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$, a function of the treatment assignment, the observed potential outcomes, and the covariates. The second is the definition of extremeness. The test statistic and definition of extremeness should be chosen to discriminate between the sharp null and alternative hypotheses. An extreme observed value of the test statistic is taken as evidence against the sharp null. As an example, we often let

$$t(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X}) = \bar{Y}_T^{\text{obs}} - \bar{Y}_C^{\text{obs}} \tag{2.1}$$

where $\bar{Y}_T^{\text{obs}}$ and $\bar{Y}_C^{\text{obs}}$ are the average observed outcomes in the treated and control groups. We can then define extremeness in terms of the absolute value of the test statistic, where larger absolute values correspond to more extreme.

After the units are assigned to treatment and control, the outcomes are recorded and the data analysis begins. In randomization tests, the observed data is re-analyzed for every treatment assignment vector in $\mathscr{S}$. More precisely, the test statistic is computed for every treatment assignment vector in $\mathscr{S}$ and the distribution of these test statistic values is called the *reference* or *null* distribution of the test statistic. The observed value is compared to this reference distribution. Let $\mathscr{S}_{\text{ref}}$ be the *reference set* of treatment assignments used to form the reference distribution. The idea of a reference set is due to Fisher (1956) and will be discussed in more detail later. At this point, we note that $\mathscr{S}_{\text{ref}} = \mathscr{S}$.

The observed potential outcomes and the sharp null hypothesis are used to impute the complete potential outcomes table and the reference distribution is created by applying every possible treatment assignment in $\mathscr{S}$ to the complete imputed potential outcomes table. Let $\mathbf{Y}^{\text{imp}} = (\mathbf{Y}^{\text{imp}}(1), \mathbf{Y}^{\text{imp}}(0))$, where $\mathbf{Y}^{\text{imp}}(1) = (Y_1^{\text{imp}}(1), \ldots, Y_N^{\text{imp}}(1))$ and $\mathbf{Y}^{\text{imp}}(0) = (Y_1^{\text{imp}}(0), \ldots, Y_N^{\text{imp}}(0))$, be the complete imputed potential outcomes table, where $\text{p}(\mathbf{Y}^{\text{imp}} \,|\, \mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathscr{M})$ gives the probability of different imputed potential outcomes tables. The distribution of $\mathbf{Y}^{\text{imp}}$ depends on $\mathbf{W}$, $\mathbf{Y}^{\text{obs}}$, $\mathbf{X}$, and the imputation model, $\mathscr{M}$. In a randomization test, the imputation model is the sharp null hypothesis, where $Y_i^{\text{imp}}(1) = Y_i^{\text{obs}}$ and $Y_i^{\text{imp}}(0) = Y_i^{\text{obs}}$. In this case, $\mathbf{Y}^{\text{imp}}$ is a deterministic function of $\mathbf{Y}^{\text{obs}}$. However, more generally, $\mathscr{M}$ can take on a variety of models and plays a key role in connecting randomization tests to posterior predictive checks (Rubin, 1984).

We lay out the randomization test in a series of five steps.

1. Select treatment assignment, $\mathbf{W} = \mathbf{w}$, from $\mathscr{S}$ and observe $\mathbf{Y}^{\text{obs}} = \mathbf{y}^{\text{obs}}$.

2. Calculate observed test statistic, $t(\mathbf{w}, \mathbf{y}^{\text{obs}}, \mathbf{X})$.

3. Using $\mathbf{w}$, $\mathbf{y}^{\text{obs}}$ and the sharp null hypothesis, impute the potential outcomes table, $\mathbf{Y}^{\text{imp}} = \mathbf{y}^{\text{imp}}$.

4. Using $\mathbf{y}^{\text{imp}}$ and $\text{p}(\mathbf{W})$, find the reference distribution

$$\text{p}(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{y}^{\text{imp}}), \mathbf{X})). \tag{2.2}$$

Using the reference distribution, find the *rejection region*, the set of "extreme"

values of the test statistic. Let $R_{\mathbf{y}^{\text{imp}}}$ be the rejection region where

$$\Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{y}^{\text{imp}}), \mathbf{X}) \in R_{\mathbf{y}^{\text{imp}}}) \leq \alpha \tag{2.3}$$

Based on the definition of extremeness, we can also define the p-value. Using the absolute value as the definition of extremeness, the p-value is

$$p = \Pr(|t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{y}^{\text{imp}}), \mathbf{X})| \geq |t(\mathbf{w}, \mathbf{y}^{\text{obs}}, \mathbf{X})|). \tag{2.4}$$

5. Reject the sharp null if $t(\mathbf{w}, \mathbf{y}^{\text{obs}}, \mathbf{X}) \in R_{\mathbf{y}^{\text{imp}}}$. Equivalently, reject the sharp null if $p \leq \alpha$.

Because $\mathscr{S}$ and $\text{p}(\mathbf{W})$ are used both to randomize the units to treatment and control and to test the sharp null hypothesis, randomization tests follow the "analyze as you randomize" principle due to Fisher (1935).

Note that $\mathscr{S}$ can be very large. For instance, in a completely randomized design, if $N = 100$ and $N_T = 50$, then $\mathscr{S} = \binom{100}{50} \approx 10^{29}$. It can be computationally prohibitive to exactly calculate the p-value (or to exactly find the region $R_{\mathbf{y}^{\text{obs}}}$). To quote Bradley (1968), randomization tests are "stunningly efficient tests that are dismally impractical." In response, Tukey (1993) defined $\mathscr{S}$ to be a much smaller set of treatment assignments, say $|\mathscr{S}| = 1000$. Then the randomization test requires minimal computational effort to compute the p-value but the validity of the test is preserved. However, today it is common to use Monte Carlo simulation (i.e. sampling from $\mathscr{S}$ according to $\text{p}(\mathbf{W})$) to approximate the p-value or $R_{\mathbf{y}^{\text{obs}}}$ and the impracticality of randomization tests no longer applies.

2.2.3 Proving randomization tests have significance level α

To prove that the randomization test has significance level α, we need to show that when the sharp null is true, the test rejects the sharp null with probability $\leq \alpha$.

Under the sharp null hypothesis, $H_0 : Y_i(1) = Y_i(0)$ for all $i = 1, \ldots, N$, the imputed potential outcomes table is equal to the true potential outcomes table, $\mathbf{y}^{\text{imp}} = \mathbf{Y}$. Thus, the distribution of the test statistic using $\mathbf{y}^{\text{imp}}$ as the potential outcomes table, $\text{p}(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{y}^{\text{imp}}), \mathbf{X}))$, is equal to the distribution of the test statistic using the true potential outcome table, $\text{p}(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{Y}), \mathbf{X}))$. Let R be the set of extreme values of the test statistic such that

$$\Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{y}^{\text{imp}}), \mathbf{X}) \in R) \leq \alpha. \tag{2.5}$$

Thus, under the sharp null, $R_{\mathbf{y}^{\text{imp}}} = R$. Finally, we can then find the unconditional probability we reject the sharp null when the sharp null is true

$$\begin{aligned} \Pr(\text{Reject } H_0) &= \Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{Y}), \mathbf{X}) \in R_{\mathbf{Y}^{\text{imp}}}) \\ &= \Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{Y}), \mathbf{X}) \in R) \\ &\leq \alpha \end{aligned} \tag{2.6}$$

where the last line follows from the definition of R. Also, note that in the first line the random variable $\mathbf{Y}^{\text{imp}}$ is in the subscript of $R_{\mathbf{Y}^{\text{imp}}}$ because we are *not* conditioning on a particular treatment assignment, $\mathbf{W} = \mathbf{w}$. Thus, the randomization test has unconditional significance level α.

2.2.4 Additive treatment effects and confidence intervals

We can generalize the derivation to show that the randomization test also has the correct significance level for testing an additive null hypothesis, $H_0 : Y_i(1) = Y_i(0) + c$ for all $i = 1, \ldots, N$. Only the imputation model, $\mathscr{M}$, changes. For instance, $Y_i(1)^{\text{imp}} = Y_i^{\text{obs}}$ if $W_i = 1$ and $Y_i(1)^{\text{imp}} = Y_i^{\text{obs}} + c$ if $W_i = 0$. Thus, step 3 of the randomization test changes but the remaining steps stay the same. The proof applies directly because under the sharp null hypothesis, it is still the case that $\mathbf{y}^{\text{imp}} = \mathbf{Y}$.

We can invert the randomization test to create a confidence interval of plausible additive effects as well. The confidence interval would consist of all c^*s such that we fail to reject the sharp null $H_0 : Y_i(1) = Y_i(0) + c^*$ for all $i = 1, \ldots, N$.

2.3 Conditional randomization tests

Cox (1982) introduced the conditional randomization test but the idea of conditional inference can be traced back to Fisher and his notion of relevant subsets. At a conceptual level, statistical inferences about a parameter θ are made by comparing the observed data to hypothetical observations that might have been observed for different values of θ. Different values of θ place different probabilities on these hypothetical observations and these probabilities lead directly to p-values and confidence intervals. For instance, if for a particular value of θ, the observed data is far less likely than some hypothetical observations, then that value of θ would not be included in the confidence interval. Fisher argued that the set of hypothetical observations used for statistical inference should not necessarily include all hypothetical observations

and should be chosen carefully. He called this set the relevant subset of hypothetical observations. To quote, Cox (1958), relevant subsets

> should be taken to consist, so far as is possible, of observations similar to the observed set in all respects which do not give a basis for discrimination between possible values of the unknown parameter of interest.

The idea of "observations similar to the observed set" is admittedly vague, and it is not immediately obvious why a subset of the hypothetical observations should lead to better inferences. The idea and its implications have been extensively studied and debated in the statistics literature. However, certain principles have become well established and we focus on those.

Relevant subsets are closely related to ancillary statistics. By definition, the distribution of ancillary statistics do not depend on the unknown parameter of interest. Also, observations with the same value of the ancillary statistic share some similarity to each other. Because the statistics do not depend on the parameter of interest, as a whole, the observations should not favor one parameter value over another. Thus, observations with the same value of the ancillary statistic form a relevant subset.

Cox (1958) gave perhaps the best known example of this idea. Say, we are interested in testing whether the temperature, μ, is less than some constant c. Let the null hypothesis be $H_0 : \mu < c$. We can use one of two unbiased thermometers. Thermometer 1 is known to be very imprecise and thermometer 2 is known to be very precise. Let $X_1 \sim N(\mu, \sigma_1^2)$ be the temperature reading from thermometer 1 and let $X_2 \sim N(\mu, \sigma_2^2)$ be the temperature reading from thermometer 2, where $\sigma_1^2 >> \sigma_2^2$. We randomly decide which thermometer to use by flipping a coin, where U is the thermometer selected and $P(U = 1) = P(U = 2) = 0.5$. The tempera-

ture reading, X, is thus drawn from a mixture of two normal distributions, where $X \sim \mathbf{1}_{U=1} \cdot X_1 + (1 - \mathbf{1}_{U=1}) \cdot X_2$ and $(U, X) = (u, x)$ is the observed data.

We consider two ways of testing the null hypothesis, $H_0 : \mu < c$. First, we can recognize that which thermometer was selected is an ancillary statistic because the distribution of U does not depend on μ. We can then only consider hypothetical observations where $U = u$. This is equivalent to conditioning on $U = u$. Thus, when $u = 1$, the test statistic is $z = \frac{x-c}{\sigma_1}$ and we reject the null hypothesis when $z > 1.68$ at the conventional 5% level. When $u = 2$, the test statistic is $z = \frac{x-c}{\sigma_2}$ and again we reject the null hypothesis when $z > 1.68$. This approach is called the conditional test and is entirely consistent with Fisher's argument. It is also intuitive. If you know you used the precise thermometer, the variance of the imprecise thermometer is irrelevant.

However, Cox (1958) showed that the conditional test is actually not the most powerful test when the alternative of interest is $\mu' \approx c + \sigma_1$. The most powerful test, called the unconditional test, is the one that rejects the null when $\frac{x-c}{\sigma_1} > 1.28$ when $u = 1$ and when $\frac{x-c}{\sigma_2} > 5$ when $u = 2$. The test increases power while holding the Type 1 error probability at 5% by trading off the conditional Type 1 error probabilities. The most powerful test has a dramatically larger Type 1 error probability when $u = 1$ compared to the conditional test. To again quote Cox (1958)

> The unconditional test says that we can assign a higher significance than we ordinarily do, because if we were to repeat the experiment, we might sample some quite different distribution. But this fact seems irrelevant to the interpretation of an observation which we know came from a distribution with variance σ_1^2. That is, our calculation of power, etc. should be made conditionally within the distribution known to have been sampled, i.e. if we are using tests of the conventional type, the conditional test should be chosen.

> To sum up, if we are to use statistical power of the conventional type, the sample space ... must not be determined solely by considerations of power, or by what would happen if the experiment were repeated indefinitely.

Birnbaum (1962) formalized this notion and called it the *conditionality principle*. The conditionality prinicple applies when an experiment, E, is a random mixture of component experiments, $E_1, \ldots, E_m$. This means that running experiment E involves, first, randomly selecting one of the component experiments and, second, running the component experiment. The conditionality principle says that the *evidential meaning* of the experiment is the same as the meaning of the randomly selected component experiment. More colloquially, "any experiment not performed is irrelevant" (Helland, 1995). That is, we can ignore the component experiments that were not selected. Cox's example fits directly into this context because the two thermometers represent two component experiments. Kalbfleisch (1975) called the selected experiment an *experimentally ancillary statistic*.

2.3.1 Conditional randomization test mechanics

Our development of the conditional randomization test parallels Kiefer (1977)'s development of the conditional confidence methodology, especially the notion of partitions. Let $\mathscr{S}_1, \ldots, \mathscr{S}_m$ partition the set of acceptable treatment assignments, $\mathscr{S}$, such that $\mathscr{S}_i \cap \mathscr{S}_j = \emptyset$ for all $i \neq j$ and $\cup_{i=1}^{m} \mathscr{S}_i = \mathscr{S}$. We shortly discuss different ways in which $\mathscr{S}_1, \ldots, \mathscr{S}_m$ are constructed, but for now, treat the partitions as given.

Let $\pi_i = \Pr(\mathbf{W} \in \mathscr{S}_i)$ be the probability of selecting a treatment assignment from the ith partition. Selecting the treatment assignment according to $\mathrm{p}(\mathbf{W})$ is equivalent to first selecting one of the partitions using the probabilities $(\pi_1, \ldots, \pi_m)$ and then

selecting the treatment assignment from the partition according to $\mathrm{p}(\mathbf{W} \mid \mathbf{W} \in \mathscr{S}_i)$, where

$$\Pr(\mathbf{W} = \omega \mid \mathbf{W} \in \mathscr{S}_i) = \frac{\Pr(\mathbf{W} = \omega)}{\sum_{\omega' \in \mathscr{S}_i} \Pr(\mathbf{W} = \omega')} \mathbf{1}_{\omega \in \mathscr{S}_i}. \tag{2.7}$$

Thus, we can frame this experiment as a mixture of component experiments, where each partition corresponds to a component experiment. Following the conditionality principle, we should only consider the selected partition of treatment assignments when carrying out the test.

In the randomization test, it was the case that $\mathscr{S}_{\text{ref}} = \mathscr{S}$. However, in a conditional randomization test, $\mathscr{S}_{\text{ref}}$ is the partition that contains the observed treatment assignment. To emphasize the fact that the reference set depends on the observed treatment assignment, we write $\mathscr{S}_{\text{ref}} = \mathscr{S}_{\text{ref}}(\mathbf{w})$. To reiterate, in a conditional randomization test, $\mathscr{S}_{\text{ref}}(\mathbf{w}) \in \{\mathscr{S}_1, \ldots, \mathscr{S}_m\}$ and $\mathbf{w} \in \mathscr{S}_{\text{ref}}(\mathbf{w})$. Consequently, the conditional randomization test does not follow the "analyze as you randomize" principle.

As we did for randomization tests, we lay out the steps of the conditional randomization test.

1. Select treatment assignment, $\mathbf{W} = \mathbf{w}$, from $\mathscr{S}$ and observe $\mathbf{Y}^{\text{obs}} = \mathbf{y}^{\text{obs}}$.

2. Calculate observed test statistic, $t(\mathbf{w}, \mathbf{y}^{\text{obs}}, \mathbf{X})$.

3. Using $\mathbf{w}$, $\mathbf{y}^{\text{obs}}$, and the sharp null hypothesis, impute the potential outcomes table, $\mathbf{Y}^{\text{imp}} = \mathbf{y}^{\text{imp}}$.

4. Using $\mathbf{y}^{\text{imp}}$ and $\mathrm{p}(\mathbf{W} \mid \mathbf{W} \in \mathscr{S}_{\text{ref}}(\mathbf{w}))$, find the reference distribution

$$\mathrm{p}(t(\mathbf{W}, \mathbf{Y}^{\mathrm{obs}}(\mathbf{W}, \mathbf{y}^{\mathrm{imp}}), \mathbf{X}) \,|\, \mathbf{W} \in \mathscr{S}_{\mathrm{ref}}(\mathbf{w})). \tag{2.8}$$

Using the reference distribution, find the rejection region, $R_{\mathbf{y}^{\mathrm{imp}}, \mathscr{S}_{\mathrm{ref}}(\mathbf{w})}$, where

$$\Pr(t(\mathbf{W}, \mathbf{Y}^{\mathrm{obs}}(\mathbf{W}, \mathbf{y}^{\mathrm{imp}}), \mathbf{X}) \in R_{\mathbf{y}^{\mathrm{imp}}, \mathscr{S}_{\mathrm{ref}}(\mathbf{w})} \,|\, \mathbf{W} \in \mathscr{S}_{\mathrm{ref}}(\mathbf{w})) \leq \alpha, \tag{2.9}$$

and the p-value, where

$$p = \Pr(|t(\mathbf{W}, \mathbf{Y}^{\mathrm{obs}}(\mathbf{W}, \mathbf{y}^{\mathrm{imp}}), \mathbf{X})| \geq |t(\mathbf{w}, \mathbf{y}^{\mathrm{obs}}, \mathbf{X})| \;|\, \mathbf{W} \in \mathscr{S}_{\mathrm{ref}}(\mathbf{w})). \tag{2.10}$$

5. Reject the sharp null if $t(\mathbf{w}, \mathbf{y}^{\mathrm{obs}}, \mathbf{X}) \in R_{\mathbf{y}^{\mathrm{imp}}, \mathscr{S}_{\mathrm{ref}}(\mathbf{w})}$. Equivalently, reject the sharp null if $p \leq \alpha$.

2.3.2 Proving conditional randomization tests have significance level α

To prove that the conditional randomization test has unconditional significance level α, we need to show that when the sharp null is true, the test rejects the sharp null with probability $\leq \alpha$.

As before, under the sharp null hypothesis, $H_0 : Y_i(1) = Y_i(0)$ for all $i = 1, \ldots, N$, the imputed potential outcomes table is equal to the true potential outcomes table, $\mathbf{y}^{\mathrm{imp}} = \mathbf{Y}$. Thus, the conditional distribution of the test statistic using $\mathbf{y}^{\mathrm{imp}}$ as the potential outcomes table, $\mathrm{p}(t(\mathbf{W}, \mathbf{Y}^{\mathrm{obs}}(\mathbf{W}, \mathbf{y}^{\mathrm{imp}}), \mathbf{X}) \,|\, \mathbf{W} \in \mathscr{S}_{\mathrm{ref}}(\mathbf{w}))$, is equal to the

conditional distribution of the test statistic using the true potential outcomes table, $\mathrm{p}(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{Y}), \mathbf{X}) \,|\, \mathbf{W} \in \mathscr{S}_{\text{ref}}(\mathbf{w}))$. Let $R_{\mathscr{S}_{\text{ref}}(\mathbf{w})}$ be the rejection region such that

$$\Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{Y}), \mathbf{X}) \in R_{\mathscr{S}_{\text{ref}}(\mathbf{w})} \,|\, \mathbf{W} \in \mathscr{S}_{\text{ref}}(\mathbf{w})) \leq \alpha. \tag{2.11}$$

Thus, under the sharp null, $R_{\mathbf{y}^{\text{imp}}, \mathscr{S}_{\text{ref}}(\mathbf{w})} = R_{\mathscr{S}_{\text{ref}}(\mathbf{w})}$. Also, for notational convenience, let $R_{\mathscr{S}_i}$ be the rejection region when $\mathbf{w} \in \mathscr{S}_i$. Thus,

$$\Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{Y}), \mathbf{X}) \in R_{\mathscr{S}_i} \,|\, \mathbf{W} \in \mathscr{S}_i) \leq \alpha \tag{2.12}$$

for all $i = 1, \ldots, m$, and $R_{\mathscr{S}_{\text{ref}}(\mathbf{w})} = R_{\mathscr{S}_i}$ if $\mathbf{w} \in \mathscr{S}_i$. We can then find the probability we reject the sharp null when the sharp null is true.

$$\begin{aligned}
\Pr(\text{Reject } H_0) &= \Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X}) \in R_{\mathbf{Y}^{\text{imp}}, \mathscr{S}_{\text{ref}}(\mathbf{W})}) \\
&= \Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X}) \in R_{\mathscr{S}_{\text{ref}}(\mathbf{W})}) \\
&= \sum_{i=1}^{m} \Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{Y}), \mathbf{X}) \in R_{\mathscr{S}_{\text{ref}}(\mathbf{W})} \,|\, \mathbf{W} \in \mathscr{S}_i) \Pr(\mathbf{W} \in \mathscr{S}_i) \\
&= \sum_{i=1}^{m} \Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{Y}), \mathbf{X}) \in R_{\mathscr{S}_i} \,|\, \mathbf{W} \in \mathscr{S}_i) \Pr(\mathbf{W} \in \mathscr{S}_i) \\
&\leq \sum_{i=1}^{m} \alpha \Pr(\mathbf{W} \in \mathscr{S}_i) \\
&= \alpha
\end{aligned} \tag{2.13}$$

In the third line of the above equation, we use the law of total probability to sum over all partitions, $\mathscr{S}_1, \ldots, \mathscr{S}_m$. Thus, the conditional randomization test has uncon-

ditional significance level α.

It is important to keep in mind that there are some restrictions on the partitions, $\mathscr{S}_1, \ldots, \mathscr{S}_m$. For a given partition, $\mathscr{S}_i$, in order for $R_{\mathscr{S}_i}$ to exist or equivalently, for the p-value to ever be $\leq \alpha$, the number of elements in $\mathscr{S}_i$ must be $\geq \alpha^{-1}$. Otherwise, even the most extreme value of the test statistic would not lead to the sharp null being rejected.

Additionally, in order for the test to have significance level α, the partitions must be specified before the experimenter has access to the observed outcomes. Otherwise, the experimenter could consciously or subconsciously manipulate the inference by changing the reference distribution. This follows Rubin (2007)'s principle of separating design from analysis.

2.3.3 Constructing partitions

When Cox (1982) introduced the idea of a conditional randomization test, he was writing about sequential clinical trials and most of the conditional randomization test literature has focused on sequential design (Wei et al., 1986; Smythe, 1988; Hollander and Peña, 1988; Mehta et al., 1988; Wei et al., 1989). Sequential clinical trials are unique in that experimental subjects enter the study serially in time and the total number of patients in the study is rarely known in advance. Additionally, experimental subjects must be assigned to treatment or control immediately upon arrival and thus, the first subject must be assigned before the second subject is assigned. Popular sequential designs include Efron's biased coin design (Efron, 1971) and the urn design (Wei, 1978). In using these designs, it is possible to end up with different numbers

of treated and control units. Cox (1982) suggested that a randomization test should be carried out conditional on "the numbers of assignments to each treatment and on any further aspects of the treatment arrangement in which there is reason to think relevant." In this setting, the partitions, $\mathscr{S}, \ldots, \mathscr{S}_m$, contain treatments assignments with the same value of N_T and N_C.

In any experiment, the experimenter will likely be faced with several different ways to form the partitions. While no general theory exists, according to Berger and Wolpert (1988), the general idea is to "find subsets ... which when conditioned upon, change the experimental measure." In our setting, this refers to the conditional properties of the test differing from the unconditional properties of the test. If the conditional and unconditional properties differ, then the subset should be conditioned on.

With that in mind, we show why conditioning on N_T makes sense. Consider the following example. An experimenter is interested in comparing two treatments and the number of experimental units is $N = 100$. The experimenter uses a Bernoulli design where experimental units are independently assigned to treatment with probability $\Pr(W_i = 1) = 0.5$. We assume the sharp null is true and the test statistic is $\hat{\tau} = \bar{Y}_T - \bar{Y}_C$. Then, in our example, the unconditional distribution of the test statistic is the black line in Figure 2.1 and the black dotted lines at -2 and 2 mark the rejection region for the unconditional test. There is a 5% chance the experimenter observes a test statisic in the rejection region.

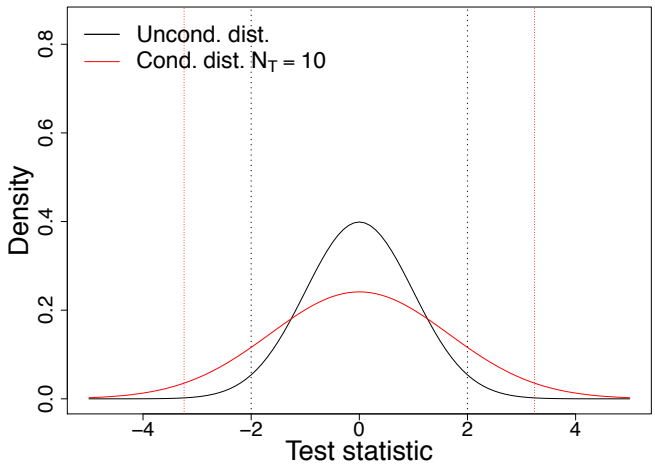


Figure 2.1: **Unconditional and conditional distributions of test statistic:** The unconditional distribution is the black solid line and the black vertical dotted lines mark the unconditional rejection region. The conditional distribution when $N_T = 10$ is the solid red line and the red vertical lines mark the conditional rejection region. The conditional probability of rejecting the test using the unconditional rejection region is 0.23.

However, say the experimenter observes an unusual treatment assignment where only 10 units are assigned to the treatment group, $N_T = 10$. At this point, even before recording the outcomes, the experimenter knows the variance of the conditional distribution of the test statistic, $\text{var}(\hat{\tau} \mid N_T = 10)$, is much larger than the variance of the unconditional distribution, $\text{var}(\hat{\tau})$. At an intuitive level, this is because $\text{var}(\hat{\tau})$ is a function of $\text{var}(\bar{Y}_T)$, $\text{var}(\bar{Y}_C)$, and a correlation term. When we condition on $N_T = 10$, $\text{var}(\bar{Y}_T \mid N_T = 10)$ increases relative to $\text{var}(\bar{Y}_T)$ because of the small number of treated units. This increase is greater than the amount $\text{var}(\bar{Y}_C \mid N_T = 10)$ decreases relative

to var($\bar{Y}_C$). We examine a similar phenomena more thoroughly in Section 2.5. That means that the probability of rejecting the sharp null is significantly higher than 0.05. In our case, the distribution of the test statistic conditional on $N_T = 10$ is the red line. The conditional probability the null is rejected is the conditional probability the test statistic is less than -2 or greater than 2. In this case, the conditional probability is 0.23. By the logic above, this implies N_T and N_C should be used to create the reference set.

The red dotted lines in Figure 2.1 mark the conditional rejection region when $N_T = 10$. Note that the conditional randomization test has the correct conditional Type I error rate. We address this point again later.

2.4 Covariate adjustment

We now return to our original question of how to adjust randomization tests for covariate imbalance. We first review the literature on two approaches. The more popular approach adjusts the randomization test by modifying the test statistic. Typically, the observed outcome is regressed on the covariates and the test statistic is defined in terms of the regression residuals. The second approach uses the conditional randomization test by conditioning on the covariate balance. We build on the conditional randomization test literature by introducing new notation to formally define covariate balance and explore the challenges regarding implementation.

2.4.1 Modifying the test statistic

A popular method of adjusting randomization tests for covariate imbalance is to first regress the observed potential outcomes on the covariates. The residuals from the regression are treated as the "adjusted outcomes" and the randomization test is carried out by calculating the test statistic using the adjusted outcomes in place of the observed potential outcomes. For instance, if Y_i^{obs} is continuous we can let the residuals be

$$e_i^{\text{obs}} = Y_i^{\text{obs}} - f(X_i) \tag{2.14}$$

where $f(\cdot)$ is a flexible, potentially non-parametric, function. The test statistic can be, for instance, the difference between the mean of the residuals in the treatment and control group,

$$t(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X}) = \bar{e}_T^{\text{obs}} - \bar{e}_C^{\text{obs}}. \tag{2.15}$$

This procedure is described in both Raz (1990) and Rosenbaum (2002). Tukey (1993) also described a similar procedure but recommended first creating "compound covariates," typically linear combinations of existing covariates, and using the compound covariates in the regression, which is similar in spirit to principal component regression. If the outcome is discrete, Gail et al. (1988) proposed using components of the score function derived from a generalized linear model as the adjusted outcome.

This approach is more robust than ANCOVA, which involves regressing $\mathbf{Y}^{\text{obs}}$ on $\mathbf{W}$ and $\mathbf{X}$ and testing the treatment effect by carrying out a t or F test for the inclusion

of **W**, because it does not assume the model is correctly specified. For instance, the nominal size for the randomization test using the residuals is maintained even when relevant covariates are not included in the regression and the assumed distribution for the outcome is incorrect. Stephens et al. (2013) carried out an extensive simulation study to compare such randomization tests to model based regression approaches, including Zhang et al. (2008)'s semi-parametric estimator. They found that the model based approaches often inflate the probability of Type I error.

2.4.2 Conditioning on covariate balance

Zheng and Zelen (2008) proposed using the conditional randomization test to analyze multi-center clinical trials by conditioning on the number of treated subjects in each center. They motivated the test primarily through simulations showing that the power of the conditional randomization test is greater than the power of the unconditional test. The emphasis on power, more precisely, unconditional power, is not surprising given that the usual rationale for covariate adjustment is increased precision and the results imply that conditioning on the observed covariate balance is similar to more traditional forms of covariance adjustment. While Zheng and Zelen (2008) only considered the multi-center clinical trial, they were confident the idea could be applied more generally.

> The novel idea of conditioning on the ancillary statistics provides an alternative method to adjust for covariates in randomization based inference ... For any arbitrary covariate, the number of patients assigned to one treatment for each level of the covariate is an ancillary statistic. Conditioning on the ancillary statistics in a randomization based analysis is a way of adjusting for the covariate effect. This idea generalizes when there are an arbitrary number of covariates. Discretized continuous covariates can also be adjusted using the same idea.

Rosenbaum (1984) also proposed a conditional randomization test but in the context of an observational study. We can view an observational study as a randomized experiment in which the treatment assignment mechanism is unknown. Thus, without knowing the probability of different assignments, we cannot generally carry out a randomization test. However, Rosenbaum (1984) identified two assumptions that enable a conditional randomization test. The two assumptions are that 1) the treatment assignment is *strongly ignorable* given $\mathbf{X}$, which means that $\mathbf{X}$ contains all the covariates that were used to make treatment assignments, and 2) the conditional probability of treatment assignment, $\mathrm{p}(\mathbf{W} \mid \mathbf{X})$, has the following form

$$\mathrm{p}(\mathbf{W} \mid \mathbf{X}) = \prod_{i=1}^{N} \pi_i^{W_i} (1 - \pi_i)^{(1-W_i)} \tag{2.16}$$

and

$$\mathrm{logit}(\pi_i) = \mathbf{X}_i \beta \tag{2.17}$$

where β is an unknown nuisance parameter and π_i is the propensity score for the ith unit (Rosenbaum and Rubin, 1983). Under the logistic regression model, $\mathbf{W}^T\mathbf{X}$ is a sufficient statistic for β. If we condition on the sufficient statistic, the conditional probability distribution of $\mathbf{W}$ will not depend on β. In fact, the conditional probability of any treatment assignment with the same value of the sufficient statistic is the same. The conditional probability of any treatment assignment with a different value of the sufficient statistic is zero.

Given this result, we can then carry out a conditional randomization test by

finding the value of the test statistic for all treatment assignments with the same value of the sufficient statistic. The sufficient statistic, $\mathbf{W}^T\mathbf{X}$, is a natural measure of covariate imbalance. For instance, if the first column of $\mathbf{X}$ is the intercept, the first element of $\mathbf{W}^T\mathbf{X}$ is the number of treated units and the remaining elements are equivalent to the covariate means in the treated group. Rosenbaum (2002) briefly noted that this method can be applied to randomized experiments. In a randomized experiment, β is in fact known, but we can still condition on its sufficient statistic. It is worth noting that the derivation does not rely on the ancillary statistic argument, but uses the fact that nuisance parameters can be eliminated by conditioning on their sufficient statistics.

Cox and Reid (2000) also addressed conditional randomization tests but viewed them as a justification for ANCOVA rather than as a practical tool. In their notation, $\bar{z}_T$ and $\bar{z}_C$ are the covariate means in the treated and control groups.

> To be relevant to the inference under discussion the ensemble of hypothetical repetitions should hold $\bar{z}_T - \bar{z}_C$ fixed, either exactly or approximately. It is possible to hold this ancillary fixed exactly only in special cases, notably when z corresponds to qualitative groupings of the units. Otherwise it can be shown that an appropriate notion of approximate conditioning induces the appropriate randomization properties for the analysis of covariance estimate ... and in that sense there is no conflict between randomization theory and that based on an assumed linear model.

2.4.3 Modified example

To illustrate why we should condition on the covariate balance, we modify our previous example as follows. As before, $N = 100$, but now assign the units assigned according to a completely randomized design where $N_T = N_C = 50$. Again, the sharp null is true and the test statistic is $\hat{\tau} = \bar{Y}_T - \bar{Y}_C$. The major difference is that now

we know whether each unit is male or female. There are 50 males and 50 females and we know that males tend to have higher potential outcomes than females. The unconditional distribution of the test statistic is the solid black line in Figure 2.2 and the black dotted lines at -2 and 2 mark the rejection region for the unconditional test. The probability the experimenter observes a test statistic in the rejection region is 0.05.

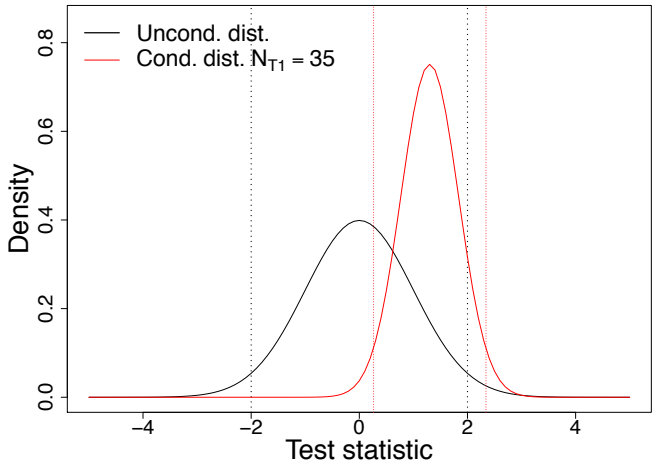


Figure 2.2: **Unconditional and conditional distributions of test statistic:** The unconditional distribution is the black solid line and the black vertical dotted lines mark the unconditional rejection region. The conditional distribution when $N_{T1} = 35$ is the solid red and the red vertical lines mark the conditional rejection region. The conditional probability of rejecting the test using the unconditional rejection region is 0.21.

The experimenter assigns units to treatment and control but ends up with an unbalanced treatment assignment. More men end up in the treated group than in the

control group. Let N_{T1} be the number of men in the treated group and N_{C1} be the number of men in the control group. In the observed treatment assignment, $N_{T1} = 35$ and $N_{C1} = 15$. This covariate imbalance creates complications for the experimenter because he knows that males and females have different potential outcome distributions. At this point, the experimenter knows that the probability of rejecting the sharp null is significantly higher than 0.05. In fact, the conditional distribution of the test statistic is the red line in Figure 2.2 and the probability of being less than -2 or greater than 2 is 0.2. By the same logic as before, this implies N_{T1} and N_{C1} should be used to create a relevant subset. The red dotted lines mark the conditional rejection region and the conditional randomization test controls the conditional Type I error rate.

2.4.4 Covariate balance function

We formalize the notion of covariate balance by introducing the covariate balance function, $B(\mathbf{W}, \mathbf{X})$, a function of $\mathbf{W}$ and $\mathbf{X}$. The covariate balance function reports a relevant summary of the covariate distribution for each level of the treatment. For instance, if the mean and variance are appropriate summaries of the covariate distribution, the covariate balance function should report the mean and variance of each covariate for each treatment level.

The value of the covariate balance function, $B(\mathbf{W}, \mathbf{X})$, is an ancillary statistic in the sense that its distribution does not depend on the treatment effect. We can then use the covariate balance function to partition the set of treatment assignments. Let $\mathscr{B}$ be the set of all possible values of covariate balance function. For each $b \in \mathscr{B}$, let

$\mathscr{S}_b = \{\omega : B(\omega, \mathbf{X}) = b\}$ be the set of treatment assignments with the same value of the covariate balance function, where $\cup_{b \in \mathscr{B}} \mathscr{S}_b = \mathscr{S}$. We carry out the conditional randomization test using these partitions.

We provide a few examples of different covariate balance functions. Consider a completely randomized design with N units where N_T are assigned to the treatment group and $N_C = N - N_T$ are assigned to the control group. In the case of one continuous covariate, $\mathbf{X} = (X_1, \ldots, X_N)$, the covariate balance function might be the mean of the covariate for each level of the treatment.

$$B(\mathbf{W}, \mathbf{X}) = \Big(\frac{1}{N_T} \sum_{i \,:\, W_i = 1} X_i, \frac{1}{N_C} \sum_{i \,:\, W_i = 0} X_i \Big) \tag{2.18}$$

Actually, because N_T, N_C, and $\sum_{i=1}^{N} X_i$ are fixed, this is equivalent to

$$B(\mathbf{W}, \mathbf{X}) = \sum_{i \,:\, W_i = 1} X_i. \tag{2.19}$$

For the case of K treatment levels, the covariate balance function would be

$$B(\mathbf{W}, \mathbf{X}) = \Big(\sum_{i \,:\, W_i = 1} X_i, \ldots, \sum_{i \,:\, W_i = K-1} X_i \Big). \tag{2.20}$$

With p continuous covariates, $\mathbf{X}_1, \ldots, \mathbf{X}_p$, and K treatment levels, the covariate balance function might be the sum of each covariate in the first $K - 1$ treatment levels.

$$B(\mathbf{W}, \mathbf{X}) = \Big(\sum_{i \,:\, W_i = 1} X_{1i}, \ldots, \sum_{i \,:\, W_i = K-1} X_{1i}, \ldots, \sum_{i \,:\, W_i = 1} X_{pi}, \ldots, \sum_{i \,:\, W_i = K-1} X_{pi} \Big) \tag{2.21}$$

We could also include the variance of the continuous covariates or define the covariate balance function on transformations of the covariates, including interactions.

One of the hazards of continuous covariates is that, to quote Rosenbaum (1984), "the number of treatment assignments ... may be too small to be of practical use." As mentioned earlier, if $|\mathscr{S}_b| < \alpha^{-1}$, the size of the conditional test will be greater than α. One possible remedy is to coarsen (i.e. round) the continuous covariates such that there are enough treatment assignments with the same covariate balance. For example, rather than report income in thousand dollar increments, report it in ten thousand dollar increments. Another approach is to discretize the continuous covariates by turning each one into a categorical covariate. In the income example, we would create income buckets, such at \$20,000-\$40,000. Both approaches destroy some information but hopefully, not too much if carried out with the help of a subject matter expert. This is reminiscent of Coarsened Exact Matching (Iacus et al., 2012), in which all covariates are discretized and balance is described by the number of units in each combination of the categorical covariates for each treatment level. Because the covariates in our motivating example are all categorical, for the purposes of this chapter, we will focus on the categorical covariate case and leave the continuous case for future work.

For categorical covariates, we can define the covariate balance function in terms of the cells of a contingency table where the rows are the levels of the covariate and the columns are the treatment levels. We start with the case of a single cateogorical covariate with J levels and a treatment with K levels, visualized in Table 2.1. Note that for a completely randomized design, the row sums and column sums are fixed.

Table 2.1: **Single categorical covariate:** For the case of one categorical covariate, the contingency table summarizes the distribution of the covariate in each level of the treatment. For a completely randomized design, a natural covariate balance function is the matrix of internal cells.

		W				
		1	2	$\cdots$	K	
X	1	$N_{1,1}$	$N_{1,2}$	$\cdots$	$N_{1,K}$	$N_{1,\cdot}$
	2	$N_{2,1}$	$N_{2,2}$	$\cdots$	$N_{2,K}$	$N_{2,\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	J	$N_{J,1}$	$N_{J,2}$	$\cdots$	$N_{J,K}$	$N_{J,\cdot}$
		$N_{\cdot,1}$	$N_{\cdot,2}$	$\cdots$	$N_{\cdot,K}$	$N_{\cdot,\cdot}$

A natural covariate balance function is the contingency table itself (i.e. the matrix of internal cells, $[N_{j,k}]$). Thus, $B(\mathbf{W}, \mathbf{X}) = [N_{j,k}]$ and if $B(\mathbf{W}, \mathbf{X}) = b$, then $\mathscr{S}_b$ is made up of those treatment assignments that produce contingency table b.

We can also use the contingency table to discuss the covariate balance function when there are multiple categorical covariates. The combinations of the categorical covariates (i.e. the Cartesian product) can be treated as the levels of a single cateogorical covariate. As an example, consider the case of two binary categorical covariates, X_1 and X_2, and a binary treatment. The contingency table considering all combinations of the covariates is shown in Table 2.2.

Table 2.2: **Multiple categorical covariates:** For the case of two categorical covariates, the combinations of the two categorical covariates can be treated as the levels of a single categorical covariate.

	W		
	0	1	
$X_1 = 0,\ X_2 = 0$	$N_{00,0}$	$N_{00,1}$	$N_{00,\cdot}$
$X_1 = 0,\ X_2 = 1$	$N_{01,0}$	$N_{01,1}$	$N_{01,\cdot}$
$X_1 = 1,\ X_2 = 0$	$N_{10,0}$	$N_{10,1}$	$N_{10,\cdot}$
$X_1 = 1,\ X_2 = 1$	$N_{11,0}$	$N_{11,1}$	$N_{11,\cdot}$
	$N_{\cdot,1}$	$N_{\cdot,2}$	$N_{\cdot,\cdot}$

In this case, we could let the covariate balance function be the contingency table. However, such a covariate balance function implies that the interaction between X_1 and X_2 is as important as X_1 and X_2 individually. While plausible in some contexts, the interaction is generally less prognostic. The number of units with $X_1 = 1$ assigned to treatment and the number of units with $X_2 = 1$ assigned to treatment are typically of greater interest. For instance, a more likely covariate balance function would be

$$B(\mathbf{W}, \mathbf{X}) = (N_{10,1} + N_{11,1}, N_{01,1} + N_{11,1}). \tag{2.22}$$

where $N_{10,1} + N_{11,1}$ are the number of units assigned to treatment with $X_1 = 1$ and $N_{01,1} + N_{11,1}$ are the number of units assigned to treatment with $X_2 = 1$. If $B(\mathbf{W}, \mathbf{X}) = b$, $\mathscr{S}_b$ consists of treatment assignments that produce the observed contingency table and treatment assignments that produce different contingeny tables consistent with $B(\mathbf{W}, \mathbf{X}) = b$. Let $\{C_{b,1}, \ldots, C_{b,l}\}$ be the set of contingency tables that satisfy $B(\mathbf{W}, \mathbf{X}) = b$. Then treatment assignments in $\mathscr{S}_b$ can be partitioned ac-

cording to their associated contingency tables. Let $\{\mathscr{C}_{b,1}, \ldots, \mathscr{C}_{b,l}\}$ be the partition of $\mathscr{S}_b$ where $\mathscr{C}_{b,i}$ is made up of the treatment assignments that produce contingency table $C_{b,i}$. We come back to this idea when we discuss sampling treatment assignments.

The covariate balance function could also make use of a cluster analysis or other methods of dimension reduction. In a cluster analysis, observations are assigned to clusters such that the observations within each cluster are more similar to each other than to those observations in other clusters. Popular clustering methods include k-means for continuous variables and k-modes for categorical variables (Huang, 1997). Clustering methods also exist for data sets with both continuous and categorical variables (Wilson and Martinez, 1997; McCane and Albert, 2008). Once the clusters have been formed, the covariates can be replaced with a single categorical covariate indicating cluster membership. The covariate balance function would then be the number of treated units within each cluster.

2.4.5 Sampling treatment assignments

In order to carry out the conditional randomization test, we must be able to draw treatment assignments from $\mathrm{p}(\mathbf{W} \mid \mathbf{W} \in \mathscr{S}_b)$. Enumerating all treatment assignment in $\mathscr{S}_b$ was explored by Rosenbaum (1984) using the backtrack algorithm. However, enumeration is only feasible for small data sets. Another approach is directly sampling treatment assignments from $\mathrm{p}(\mathbf{W})$ and accepting the treatment assignment if it is in $\mathscr{S}_b$ (i.e. rejection sampling). This too is only feasible for small data sets. For larger data sets, it is always possible to sample from $\mathrm{p}(\mathbf{W} \mid \mathbf{W} \in \mathscr{S}_b)$ in the case of a single categorial covariate. Unfortunately, more general sampling methods do not exist.

However, we discuss sampling methods for related problems that might serve as a starting point for future work. We also discuss the role of approximate conditioning.

We reiterate the importance of separating design from analysis, which implies that the experimenter should specify the covariate balance function and sampling method before having access to the observed outcomes.

Sampling from $\mathbf{p}(\mathbf{W} \,|\, \mathbf{W} \in \mathscr{S}_b)$

We return to the case of a completely randomized design with one categorical covariate with J levels and K treatment levels, represented in Table 2.1. In a completely randomized design, all treatment assignments are equally likely, which implies that $\mathrm{p}(\mathbf{W} \,|\, \mathbf{W} \in \mathscr{S}_b) = |\mathscr{S}_b|^{-1}$. As before, let the covariate balance function be the number of units at each level of the categorical covariate assigned to each treatment level. The goal is to sample treatment assignments such that the internal cells in Table 2.1 are held fixed. This is easily accomplished by independently permuting the treatment assignments of the units at each level of the categorical covariate (i.e. in each row). In this way, the values of the internal cells remain constant. This method can be applied to generate treatment assignments from any contingency table. Thus, whenever $\mathscr{S}_b$ is made up of treatment assignments that produce the same contingency table, sampling treatment assignments is straightforward.

Unfortunately, when $\mathscr{S}_b$ is made up of treatment assignments associated with different contingency tables, sampling treatment assignments is significantly more complicated. Ideally, we could proceed in two steps. First, sample a contingency table $C_{b,i}$ from $\{C_{b,1}, \ldots, C_{b,l}\}$, where the probability of selecting $C_{b,i}$ is $\frac{|\mathscr{C}_{b,i}|}{|\mathscr{S}_b|}$. Second,

sample a treatment assignment from $\mathrm{p}(\mathbf{W} \mid \mathbf{W} \in \mathscr{C}_{b,i})$. We just showed that the second step is straightforward but the challenge is sampling from the set of contingency tables $\{C_{b,1}, \ldots, C_{b,l}\}$. Remember that $\{C_{b,1}, \ldots, C_{b,l}\}$ is a set of contingency tables with fixed marginals that satisfy additional constaints on certain cells. How to sample contingency tables with fixed marginals has been extensively studied (Mehta and Patel, 1983), (Diaconis and Sturmfels, 1998), (Chen et al., 2005), and (Chen et al., 2006) and those methods could provide a starting point for our problem. At this point, we leave further examinination for future work but note that the Markov chain Monte Carlo (MCMC) method proposed in Diaconis and Sturmfels (1998) could be useful in sampling contingency tables that satisfy $B(\mathbf{W}, \mathbf{X}) = b$. However, designing the Markov moves such that the constraints are satisfied can be difficult and the chain may fail to be irreducibile.

Approximate conditioning

While sampling treatment assignments from $\mathscr{S}_b$ is only feasible in the simplest case, it is often possible to sample from either a subset or superset of $\mathscr{S}_b$. Cox (1984) proposed letting the reference set be a superset of $\mathscr{S}_b$.

> One way of ameliorating the effect of discreteness useful in extreme cases is by approximate conditioning, i.e. by carefully assembling conditional distributions given ancillary values close to that observed.

We first discuss letting the reference set be a superset of $\mathscr{S}_b$. For each value of b, let $\gamma(b) \subset \mathscr{B}$ be the set of values of the covariate balance function close to b. Instead of sampling from $\mathrm{p}(\mathbf{W} \mid \mathbf{W} \in \mathscr{S}_b)$, we sample from $\mathrm{p}(\mathbf{W} \mid \mathbf{W} \in \cup_{b' \in \gamma(b)} \mathscr{S}_{b'})$. For instance, with continuous covariates, we could sample treatment assignments such

that $b_{\text{lb}} < \sum_{i\,:\,W_i=1} X_i < b_{\text{ub}}$. This can potentially increase the number of treatment assignments enough that we can use rejection sampling to sample treatment assignments. The same idea can be applied to categorical covariates. The disadvantage of approximate conditioning is that there is no guarantee that the test still has significance level α. To see this, note that the rejection region is now $R_{\gamma(b)}$ and $R_{\gamma(b)}$ is formed using the conditional distribution $\mathrm{p}(\mathbf{W} \,|\, \mathbf{W} \in \cup_{b' \in \gamma(b)} \mathscr{S}_{b'})$. This means that

$$\Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{Y}), \mathbf{X}) \in R_{\gamma(b)} \,|\, \mathbf{W} \in \cup_{b' \in \gamma(b)} \mathscr{S}_{b'}) \leq \alpha. \tag{2.23}$$

However, this does not imply that

$$\Pr(t(\mathbf{W}, \mathbf{Y}^{\text{obs}}(\mathbf{W}, \mathbf{Y}), \mathbf{X}) \in R_{\gamma(b)} \,|\, \mathbf{W} \in \mathscr{S}_b) \leq \alpha, \tag{2.24}$$

which is necessary in order that the test has significance level α. However, the necessary condition will hold if $R_{\gamma(b)} = R_b$. It is unlikely that $R_{\gamma(b)} = R_b$ exactly, but, as long as the set $\gamma(b)$ does not include values too dissimilar from b, it is reasonable to think that $R_{\gamma(b)} \approx R_b$. Additionally, we can assess whether $R_{\gamma(b)} \approx R_b$ by monitoring how $R_{\gamma(b)}$ changes for different specifications of $\gamma(\cdot)$. If the the rejection region is relatively constant, we can feel more confident that the test has significance level approximately α.

The story is the same when the reference set is a subset of $\mathscr{S}_b$. When all the covariates are categorical, we can consider the contingency table based on the combination of categorical covariates. We can then let the reference set consist of the treatment assignments that produce the observed contingency table since we can easily sample

those treatment assignments. Similarly, if we were able to obtain a subset of the contingency tables that satisfy $B(\mathbf{W}, \mathbf{X}) = b$, we could let the reference set be those treatment assignments that produce one of the contingency tables in the subset.

2.4.6 Rerandomization

The conditional randomization test conditioning on the observed covariate balance shares similarities with rerandomization (Morgan and Rubin, 2012). Rerandomization is a treatment assignment mechanism that restricts $\mathscr{S}$ to the set of treatment assignments which satisfy a pre-determined level of covariate balance. A balance criterion, $\phi(\mathbf{W}, \mathbf{X})$, determines if the treatment assignment is acceptable, $\phi(\mathbf{W}, \mathbf{X}) = 1$, or unacceptable, $\phi(\mathbf{W}, \mathbf{X}) = 0$. Thus, $\mathscr{S} = \{\omega : \phi(\omega, \mathbf{X}) = 1\}$. As a result, the observed treatment assignment is guaranteed to be balanced on covariates. The experiment is analyzed using a randomization test where the reference set is $\mathscr{S}$. Rerandomization is similar to restriced randomization but what makes rerandomization unique is that it is designed to balance multiple covariates simultaneously.

The conditional randomization test is like a *post-hoc rerandomization test*. In a conditional randomization test, we observe some treatment assignment, $\mathbf{W} = \mathbf{w}$, and covariate balance, $B(\mathbf{w}, \mathbf{X}) = b$, and then act as if that treatment assignment were drawn from some partition with the same covariate balance, $\mathscr{S}_b$. The rerandomization test and conditional randomization test would be identical if, for instance, $\mathscr{S}_b = \{\omega : \phi(\omega, \mathbf{X}) = 1\}$.

2.5 Post-stratification simulation

We carry out a simulation study to evaluate the unconditional and conditional properties of the conditional randomization test and compare it to the unconditional randomization test. For this simulation, the relevant unconditional properties of the tests are the average rejection rates over repeated runs of the experiment. The conditional properties of the test are the average rejection rates under repeated runs of the experiment where the covariate balance is held fixed. For a given experiment, the conditional rejection rates are arguably more relevant than the unconditional rejection rates. While the unconditional rejection rates measure the performance of the test over all treatment assignments, the conditional rejection rates measure the performance of the test for treatment assignments like the observed one. To again quote Cox (1958),

> Our calculation of power, etc. should be made conditionally within the distribution known to have been sampled.

We consider the case of a single categorical covariate. Adjusting for such a covariate is often called post-stratifcation and we refer to the levels of the covariate as strata. Pattanayak (2011) and Miratrix et al. (2013) studied post-stratification from the Neymanian perspective and derived the unconditional and conditional distributions of two estimators. We utilize many of their results in evaluating the properties of unconditional and conditional randomization tests.

Let $\mathbf{X} = (X_1, \ldots, X_N)$ be the vector of strata indicators and let J be the number of strata. The units in strata j are given by $\nu_j = \{i : X_i = j\}$. We assume a completely randomized design with two treatment levels. N_T units are assigned to

treatment and $N_C = N - N_T$ units are assigned to control. We define the covariate balance function to be the number of treated units in the each stratum,

$$\begin{aligned} B(\mathbf{W}, \mathbf{X}) &= \Big(\sum_{i \in \nu_1} W_i, \ldots, \sum_{i \in \nu_J} W_i \Big) \\ &= (N_{T1}, \ldots, N_{TJ}). \end{aligned} \tag{2.25}$$

Here, N_{Tj} is the number of treated units in the jth stratum and N_j is the number of units in the jth statum. Note that $N_T = \sum_{j=1}^{J} N_{Tj}$ and $N_C = \sum_{j=1}^{J} N_{Cj}$. Let $\bar{Y}_j^{\text{obs}}$ be the observed mean outcome in the jth stratum,

$$\bar{Y}_j^{\text{obs}} = \frac{1}{N_j} \sum_{i \in \nu_j} Y_i^{\text{obs}} \tag{2.26}$$

and $\bar{Y}^{\text{obs}} = \frac{1}{N} \sum_{j=1}^{J} N_j \bar{Y}_j^{\text{obs}}$.

2.5.1 Test statistics

In the context of post-stratificaiton, there are two primary test statistics. The first is the simple difference between treated and control means,

$$\begin{aligned} \hat{\tau}_{sd} &= \frac{1}{N_T} \sum_{i=1}^{N} W_i Y_i^{\text{obs}} - \frac{1}{N_C} \sum_{i=1}^{N} (1 - W_i) Y_i^{\text{obs}} \\ &= \bar{Y}_T^{\text{obs}} - \bar{Y}_C^{\text{obs}}. \end{aligned} \tag{2.27}$$

The second is the post-stratified test statistic, $\hat{\tau}_{ps}$, where $\hat{\tau}_{ps}$ is a weighted average of within stratum simple differences. It is defined through the strata level estimates, $\hat{\tau}_{sd,j}$, where

$$\begin{aligned}\hat{\tau}_{sd,j} &= \frac{1}{N_{Tj}} \sum_{i \in \nu_j} W_i Y_i^{\text{obs}} - \frac{1}{N_{Cj}} \sum_{i \in \nu_j} (1 - W_i) Y_i^{\text{obs}} \\ &= \bar{Y}_{Tj}^{\text{obs}} - \bar{Y}_{Cj}^{\text{obs}}. \end{aligned} \tag{2.28}$$

Then $\hat{\tau}_{ps}$ is defined to be

$$\hat{\tau}_{ps} = \sum_{j=1}^{J} \frac{N_k}{N} \hat{\tau}_{sd,j}. \tag{2.29}$$

While $\hat{\tau}_{sd}$ ignores the covariate information, $\hat{\tau}_{ps}$ utilizes it. We reviewed earlier that randomization tests can be adjusted for covariate imbalance through the test statistic and the post-stratified test statistic is an example. In fact, we can show that the post-stratified test statistic is unchanged if we replace the observed outcomes with the residuals from regressing the observed outcomes on the covariate. In the case of a single categorical covariate, the regression residual,

$$e_i^{\text{obs}} = Y_i^{\text{obs}} - f(X_i), \tag{2.30}$$

takes a simple form. If $i \in \nu_j$, then $f(X_i) = \bar{Y}_j^{\text{obs}}$. The residual based post-stratified test statistic is $\hat{\tau}_{ps}^{\text{res}}$, where the within statum differences are

$$\hat{\tau}^{\text{res}}_{sd,j} = \frac{1}{N_{Tj}} \sum_{i \in \nu_j} W_i e_i^{\text{obs}} - \frac{1}{N_{Cj}} \sum_{i \in \nu_j} (1 - W_i) e_i^{\text{obs}}. \tag{2.31}$$

and then

$$\hat{\tau}^{\text{res}}_{ps} = \sum_{j=1}^{K} \frac{N_j}{N} \hat{\tau}^{\text{res}}_{sd,j}. \tag{2.32}$$

However, it turns out that $\hat{\tau}^{\text{res}}_{sd,j} = \hat{\tau}_{sd,j}$ and thus, $\hat{\tau}^{\text{res}}_{ps} = \hat{\tau}_{ps}$. Note that

$$\begin{aligned}
\hat{\tau}^{\text{res}}_{sd,j} &= \frac{1}{N_{Tj}} \sum_{i \in \nu_j} W_i e_i^{\text{obs}} - \frac{1}{N_{Cj}} \sum_{i \in \nu_j} (1 - W_i) e_i^{\text{obs}} \\
&= \frac{1}{N_{Tj}} \sum_{i \in \nu_j} W_i (Y_i^{\text{obs}} - \bar{Y}_j^{\text{obs}}) - \frac{1}{N_{Cj}} \sum_{i \in \nu_j} (1 - W_i)(Y_i^{\text{obs}} - \bar{Y}_j^{\text{obs}}) \\
&= \frac{1}{N_{Tj}} \sum_{i \in \nu_j} W_i Y_i^{\text{obs}} - \frac{1}{N_{Cj}} \sum_{i \in \nu_j} (1 - W_i) Y_i^{\text{obs}} \\
&- \frac{1}{N_{Tj}} \sum_{i \in \nu_j} W_i \bar{Y}_j^{\text{obs}} + \frac{1}{N_{Cj}} \sum_{i \in \nu_j} (1 - W_i) \bar{Y}_j^{\text{obs}} \\
&= \hat{\tau}_{sd,j} - \frac{N_{Tj}}{N_{Tj}} \bar{Y}_j^{\text{obs}} + \frac{N_{Cj}}{N_{Cj}} \bar{Y}_j^{\text{obs}} \\
&= \hat{\tau}_{sd,j}
\end{aligned} \tag{2.33}$$

This implies that the post-stratified test statistic, $\hat{\tau}_{ps}$, adjusts for covariate imbalance.

Conditional equivalence

Another interesting result regarding $\hat{\tau}_{sd}$ and $\hat{\tau}_{ps}$ is that the conditional randomization tests using $\hat{\tau}_{sd}$ and $\hat{\tau}_{ps}$ are in fact equivalent. Thus, when conditioning on

the covariate balance, $(N_{T1}, \ldots, N_{TJ})$, the test statistic that utilizes the covariate information performs no better than the test statistic that ignores it. Similarly, this implies that conditioning on the covariate balance adjustes for covariate imbalance. We can prove the result by adapting a proof from Rosenbaum (1984). We show that the conditional randomization tests using $\hat{\tau}_{sd}$ and $\hat{\tau}_{ps}$ are equivalent by showing that, conditionally, $\hat{\tau}_{ps}$ is monotonic function of $\hat{\tau}_{sd}$.

Note that $\hat{\tau}_{ps} = \hat{\beta}_W$, where $\hat{\beta}_W$ is the estimate of β_W from the linear regression

$$Y_i^{\text{obs}} = \beta_0 + \beta_W W_i + \sum_{k=2}^{K} \beta_k X_{ik} + \sum_{k=2}^{K} \gamma_k (W_i \cdot X_{ik}) + \epsilon_i \tag{2.34}$$

where

$$X_i = \begin{cases} 1 & : \text{if the } i\text{th unit is in the } k\text{th stratum} \\ -1 & : \text{if the } i\text{th unit is in the first stratum} \\ 0 & : \text{otherwise.} \end{cases} \tag{2.35}$$

Note that X_i follows the sum contrast coding. The next step is to show that, conditioning on the observed balance, $\hat{\tau}_{ps}$ is a monotonic function of $\hat{\tau}_{sd}$.

Let $[\mathbf{W}, \mathbf{F}]$ denote the design matrix, where $\mathbf{F}$ includes a column of ones and columns for the categorical indicator variables and interactions. Also, note that $\mathbf{W}^T\mathbf{Y}^{\text{obs}} = (\hat{\tau}_{sd} + \frac{1}{N_C}\mathbf{1}^T\mathbf{Y}^{\text{obs}})/(\frac{1}{N_T} + \frac{1}{N_C})$. Let $P_{\mathbf{F}} = \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T$ be the projection matrix onto the columns of $\mathbf{F}$. We then use the regression anatomy formula (**?**).

$$\hat{\tau}_{ps} = \hat{\beta}_W = \frac{\mathbf{W}^T(I - P_{\mathbf{F}})\mathbf{Y}^{\text{obs}}}{\mathbf{W}^T(I - P_{\mathbf{F}})\mathbf{W}}.$$

Note that conditioning on the observed balance implies that $\mathbf{W}^T\mathbf{F}$ is a constant and thus

$$\hat{\tau}_{ps} = \frac{\mathbf{W}^T\mathbf{Y}^{\text{obs}} - k_1}{k_2}$$

where $k_1 = \mathbf{W}^T P_{\mathbf{F}} \mathbf{Y}^{\text{obs}}$ and $k_2 = \mathbf{W}^T(I - P_{\mathbf{F}})\mathbf{W}$. Finally, since $\mathbf{W}^T\mathbf{Y}^{\text{obs}}$ is a monotonic function of $\hat{\tau}_{sd}$, $\hat{\tau}_{ps}$ is also a monotonic function of $\hat{\tau}_{sd}$.

Mean and variance

Before we interpret the results of the simulations, it will be helpful to review the unconditional and conditional mean and variance of the two test statitics under the sharp null hypothesis. These results are special cases of the more general results derived in Pattanayak (2011) and Miratrix et al. (2013).

All means and variances are reported under the sharp null hypothesis, $H_0 : Y_i(1) = Y_i(0)$ for all $i = 1, \ldots, N$. For notational convenience, we let $Y_i = Y_i(1) = Y_i(0)$, $\bar{Y}_j = \bar{Y}_j(1) = \bar{Y}_j(0)$, and $\bar{Y} = \bar{Y}(1) = \bar{Y}(0)$. The unconditional mean and variance for $\hat{\tau}_{sd}$ are

$$\begin{aligned} E(\hat{\tau}_{sd}) &= 0 \\ \text{var}(\hat{\tau}_{sd}) &= \frac{Ns^2}{N_T N_C}, \end{aligned} \tag{2.36}$$

where $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$. The unconditional mean and variance for $\hat{\tau}_{ps}$ are

$$
\begin{aligned}
E(\hat{\tau}_{ps}) &= 0 \\
\text{var}(\hat{\tau}_{ps}) &= \sum_{j=1}^{J} s_j^2 \frac{N_j}{N}\Big(2 + \Big(N_T N_C + \frac{N^2}{N_j} - N\Big)\Big(\frac{1}{N_T^2} + \frac{1}{N_C^2}\Big)\Big),
\end{aligned} \tag{2.37}
$$

where $s_j^2 = \frac{1}{N-1}\sum_{i \in \nu_j}^{N_j}(Y_i - \bar{Y})^2$. The two test statistics have the same expected value but the variances do differ. Note that $\text{var}(\hat{\tau}_{sd})$ depends on s^2, where $(N-1)s^2$ can be thought of as the total sum of squares (SST). $\text{var}(\hat{\tau}_{ps})$ depends on s_j^2, $j = 1, \ldots, J$, where $(N_j - 1)s_j^2$ are the within stratum sum of squares (SSW). Remember that SST and SSW are connected by the familiar decomposition,

$$
\begin{aligned}
SST &= SSB + SSW \\
(N-1)s^2 &= \sum_{j=1}^{J} N_j(\bar{Y}_j - \bar{Y})^2 + \sum_{j=1}^{J}(N_j - 1)s_j^2.
\end{aligned} \tag{2.38}
$$

The more prognostic the strata, the larger $SSB = \sum_{j=1}^{J} N_j(\bar{Y}_j - \bar{Y})^2$ and the smaller the within statum sum of squares. As the strata become more prognostic, $\text{var}(\hat{\tau}_{ps})$ decreases but if SST is constant, $\text{var}(\hat{\tau}_{sd})$ stays the same.

Conditional on $(N_{T1}, \ldots, N_{TJ})$, the mean and variance for $\hat{\tau}_{sd}$ are

$$
\begin{aligned}
E(\hat{\tau}_{sd} \,|\, N_{T1}, \ldots, N_{TJ}) &= \frac{1}{N_T} \sum_{j=1}^{J} N_{Tj} \bar{Y}_j - \frac{1}{N_C} \sum_{j=1}^{J} N_{Cj} \bar{Y}_j \\
\text{var}(\hat{\tau}_{sd} \,|\, N_{T1}, \ldots, N_{TJ}) &= \Big(\frac{N}{N_T N_C} \Big)^2 \sum_{j=1}^{J} s_j^2 N_j \Big(\frac{N_{Tj}}{N_j} \Big) \Big(1 - \frac{N_{Tj}}{N_j} \Big).
\end{aligned} \tag{2.39}
$$

Note that, in general, $\hat{\tau}_{sd}$ is conditionally biased. For instance, if only the units in strata j have large potential outcomes and more units in strata j are assigned to the treatment group, $E(\hat{\tau}_{sd} \,|\, N_{T1}, \ldots, N_{TJ})$ will be positive. However, $\hat{\tau}_{sd}$ is conditionally unbiased when either $\bar{Y}_j = \bar{Y}$ or $\frac{N_{Tj}}{N_T} = \frac{N_{Cj}}{N_C}$. The first condition implies that the strata are not prognostic and the second condition implies that the treatment assignment is perfectly balanced. Regarding the conditional variance, note that $\Big(\frac{N_{Tj}}{N_j} \Big) \Big(1 - \frac{N_{Tj}}{N_j} \Big)$ is largest when $\frac{N_{Tj}}{N_j} = 0.5$. As we move further from 0.5, the conditional variance decreases. We can think of $\Big(\frac{N_{Tj}}{N_j} \Big) \Big(1 - \frac{N_{Tj}}{N_j} \Big)$ as inversely related to covariate imbalance. The smaller $\Big(\frac{N_{Tj}}{N_j} \Big) \Big(1 - \frac{N_{Tj}}{N_j} \Big)$, the bigger the covariate imbalance.

For $\hat{\tau}_{ps}$, the conditional expectation and variance are

$$
\begin{aligned}
E(\hat{\tau}_{ps} \,|\, N_{T1}, \ldots, N_{TJ}) &= 0 \\
\text{var}(\hat{\tau}_{ps} \,|\, N_{T1}, \ldots, N_{TJ}) &= \sum_{j=1}^{J} s_j^2 \frac{N_j^2}{N^2} \Big(\frac{1}{N_{Tj}} + \frac{1}{N_{Cj}} \Big) \\
&= \sum_{j=1}^{J} s_j^2 \frac{N_j^2}{N^2} \Big[\Big(\frac{N_{Tj}}{N_j} \Big) \Big(1 - \frac{N_{Tj}}{N_j} \Big) \Big]^{-1}.
\end{aligned} \tag{2.40}
$$

Again, the term $\Big(\frac{N_{Tj}}{N_j} \Big) \Big(1 - \frac{N_{Tj}}{N_j} \Big)$ appears in the conditional variance. However, for $\hat{\tau}_{ps}$,

as the covariate imbalance increases, $\left(\frac{N_{Tj}}{N_j}\right)\left(1-\frac{N_{Tj}}{N_j}\right)$ decreases, and the conditional variance increases.

While these means and variances help us interpret the simulation results, they might be used to obtain asymptotic results for the unconditional and conditional properties of the tests. Following Ding (2014), we could potentially apply the finite population central limit theorem but we leave this for future work.

2.5.2 Simulation set-up

The goal of the simulation study is to evaluate the unconditional and conditional properties of the conditional randomization test in a simple post-stratification setting. We compare the conditional and unconditional randomization tests over several simulation settings and both test statistics. We let $N = 100$, $N_T = 50$, and $N_C = N - N_T = 50$. We also let the number of strata be $J = 2$ and $\nu_1 = \nu_2 = 50$.

Table 2.3: **Simulation design:** We use a completely randomized design where $N = 100$ and $N_T = 50$.

		W		
		1	0	
X	1	N_{T1}	N_{C1}	$N_1 = 50$
	2	N_{T2}	N_{C2}	$N_2 = 50$
		$N_T = 50$	$N_C = 50$	$N = 100$

Because there are only two strata and two treatment levels, the covariate balance function is completely determined by the top left cell, N_{T1}, in Table 2.3.

We generate the "science", the complete potential outcomes table, by varying two

parameters, τ and λ. Here, τ is the additive treatment effect and λ controls the association between X and $Y(0)$ (i.e. the prognostic effect).

$$\begin{aligned} \tau &= Y_i(1) - Y_i(0) \\ \lambda &= E(Y(0) \,|\, X = 2) - E(Y(0) \,|\, X = 1) \end{aligned} \tag{2.41}$$

We let τ take on one of 11 values, $\tau \in \{0, 0.1, 0.2, \ldots, 1\}$ and λ take on one of three values, $\lambda \in \{0, 1.5, 3\}$. We generate the complete potential outcomes by first drawing $Y_i(0) \,|\, X_i$ and then filling in $Y_i(1)$ as follows.

$$\begin{aligned} Y_i(0) \,|\, X_i &\sim N(\lambda X_i, 1) \\ Y_i(1) &= Y_i(0) + \tau \end{aligned} \tag{2.42}$$

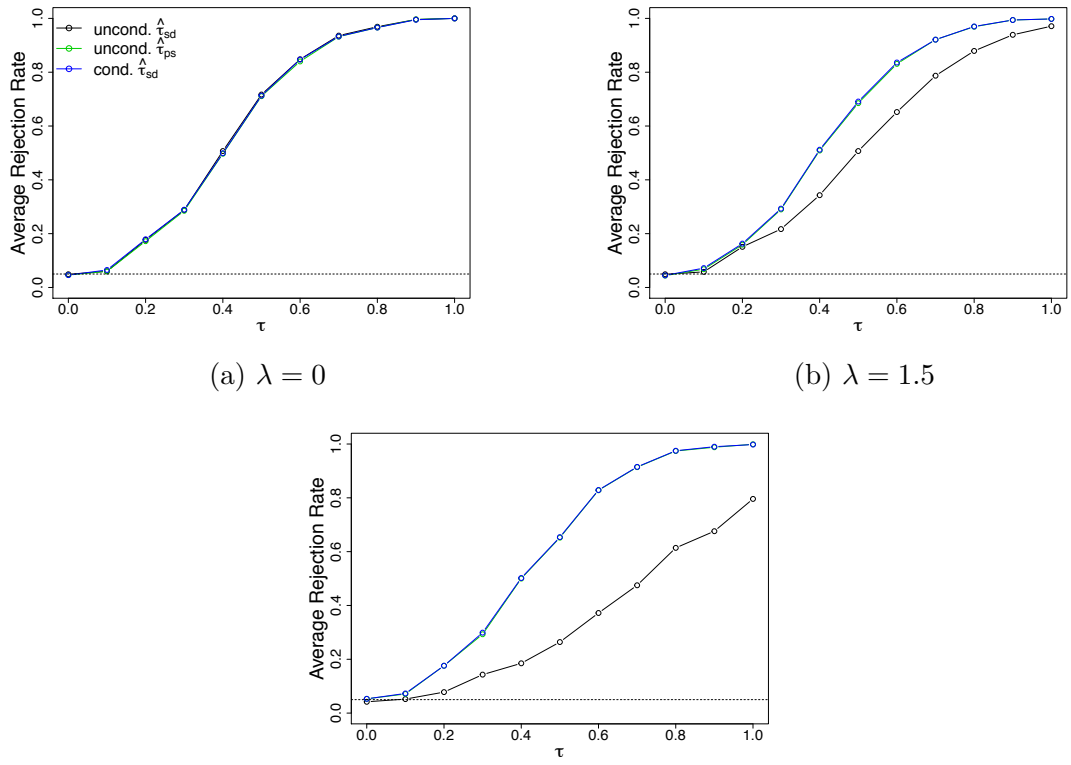
After generating the potential outcomes, we randomly assign units to treatment and control and record whether each of the three tests (two unconditional tests and one conditional test) rejects the sharp null, $H_0 : Y_i(1) = Y_i(0)$ for $i = 1, \ldots, N$, at the 0.05 significance level. We repeat this 1000 times and record the average rejection rate for each test.

We randomly assign the units in one of two ways. We either assign them using the completely randomized assignment mechanism or we assign them holding N_{T1} fixed at either 25, 30, 35, or 40. Assigning the units using the completely randomized assignment mechanism allows us to evaluate the unconditional properties of the test

and holding N_{T1} fixed allows us to assess the conditional properties of the test (i.e. how the test performs for particular values of N_{T1}). Since we are implicitly interested in situations where the covariate is prognostic, when evaluating the conditional properties, we let $\lambda = 3$.

2.5.3 Unconditional properties

Figure 2.3 reports the unconditional rejection rates for different values of τ and λ. The units were assigned using the completely randomized assignment mechanism.



(a) $\lambda = 0$

(b) $\lambda = 1.5$

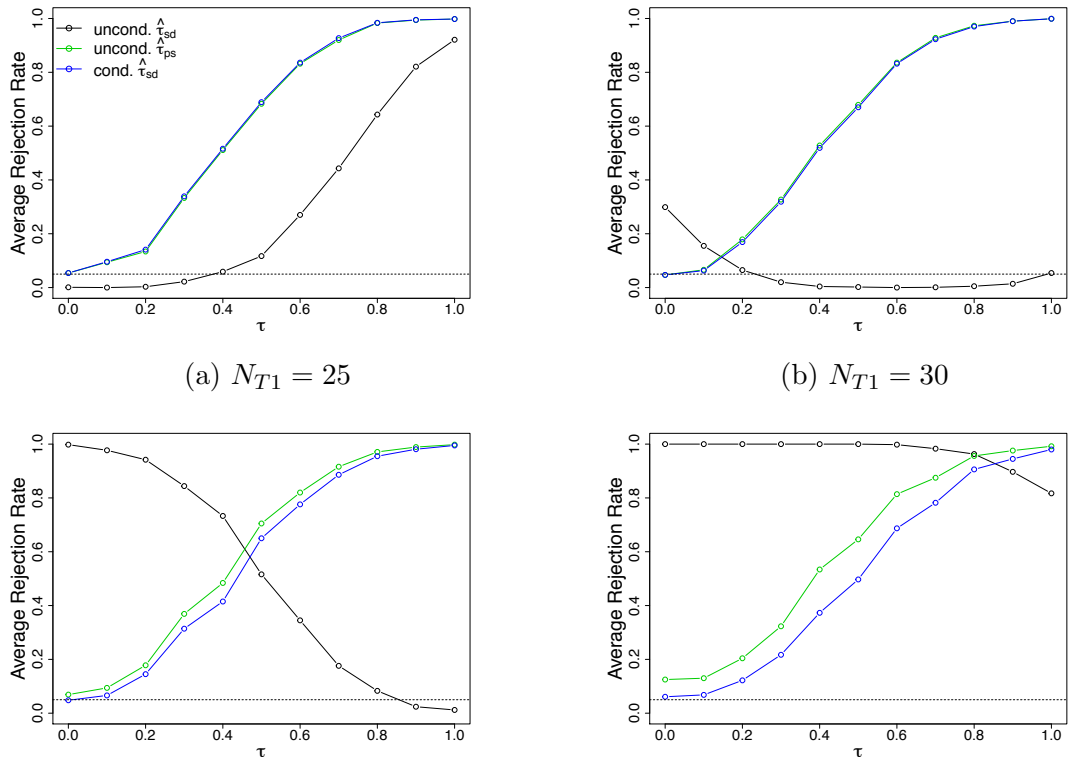
(c) $\lambda = 3$

Figure 2.3: **Unconditional average rejection rates for different τ and λ**

When $\lambda = 0$, Figure 2.3(a), the covariate is not prognostic and the three tests are virtually the same. All reject the null hypothesis with probability 0.05 (the horizontal dotted line) when the null is true, $\tau = 0$, and, as expected, the power increases as τ increases. In Figure 2.3(b), the covariate is more prognostic, $\lambda = 1.5$, and the unconditional test using $\hat{\tau}_{ps}$ and the conditional test appear unchanged but the power of the unconditional test using $\hat{\tau}_{sd}$, shown in the black line, falls. The unconditional test using $\hat{\tau}_{sd}$ is the one test that ignores the covariate balance. It is more of the same in Figure 2.3(c), where again the unconditional test using $\hat{\tau}_{ps}$ and the conditional test appear unchanged. However, the power of the unconditional test using $\hat{\tau}_{sd}$ falls even lower. In summary, as the covariate becomes more prognostic, the power of the unconditional test using $\hat{\tau}_{sd}$ decreases while the power of the other two tests remain the same. The unconditional properties support the notion that we should adjust for covariate imbalance either by modifying the test statistic or by conditioning but do not distinguish between the two approaches.

2.5.4 Conditional properties

Figure 2.4 reports the conditional rejection rates for the three tests, varying the values of τ and N_{T1}, where $\lambda = 3$.



(a) $N_{T1} = 25$ (b) $N_{T1} = 30$

(c) $N_{T1} = 35$ (d) $N_{T1} = 40$

Figure 2.4: **Conditional average rejection rates for different τ and N_{T1}:** In all simulations, $\lambda = 3$.

When $N_{T1} = 25$, Figure 2.4(a), the prognostic covariate is perfectly balanced. When $\tau = 0$, both the unconditional test using $\hat{\tau}_{ps}$ and the conditional test reject the sharp null with probability 0.05. The unconditional test using $\hat{\tau}_{sd}$ rejects the sharp null with probability less than 0.05. Remember that $E(\hat{\tau}_{sd}) = 0$ and because the covariate is perfectly balanced, $E(\hat{\tau}_{sd} \,|\, N_{T1} = 25) = 0$. But, because the covariate is prognostic, $\text{var}(\hat{\tau}_{sd}) > \text{var}(\hat{\tau}_{sd} \,|\, N_{T1} = 25)$. The unconditional randomization test using $\hat{\tau}_{sd}$ compares the test statistic to a reference distribution centered at 0 and

with variance var($\hat{\tau}_{sd}$); however, when $N_{T1} = 25$, the observed test statistics have a smaller variance. Thus, the test statistics rarely end up in the tails of the reference distribution and the rejection rate is less than 0.05.

As we move from perfect covariate balance to covariate imbalance, Figure 2.4(b), the unconditional test using $\hat{\tau}_{ps}$ and the conditional test appear unchanged, but the unconditional test using $\hat{\tau}_{sd}$ begins to break down. When $\tau = 0$, $E(\hat{\tau}_{sd}) = 0$ but because the covariate is prognostic, $E(\hat{\tau}_{sd} \,|\, N_{T1} = 30) < 0$. Thus, the unconditional test is comparing the observed test statistics, which tend to be negative, to a reference distribution centered at 0. As seen in Figure 2.4(b), this implies that the rejection rate should be greater than 0.05 when $\tau = 0$. As τ increases, $E(\hat{\tau}_{sd} \,|\, N_{T1} = 30)$ increases since the positive treatment effect counteracts the effect of the covariate imbalance. Thus, the observed test statistics are pushed closer to 0 and the rejection rate falls. Eventually, the treatment effect overcomes the covariate imbalance and the rejection rate begins to rise, which we see at $\tau = 1$.

In Figures 2.4(c) and 2.4(d), as the covariate imbalance increases, the unconditional test using $\hat{\tau}_{sd}$ repeats this pattern. More interestingly, as the covariate imbalance increases, we begin to see differences between the unconditional test using $\hat{\tau}_{ps}$ and the conditional test. Note that in Figure 2.4(d), the unconditional test using $\hat{\tau}_{ps}$ rejects the sharp null with probability over 0.05 when $\tau = 0$. This implies that the test has the wrong conditional significance level. In contrast, although the power of the conditional test has dropped slightly, its conditional significance level is still 0.05. The key to understanding why the conditional significance level is incorrect for the unconditional test using $\hat{\tau}_{ps}$ is that the conditional variance of $\hat{\tau}_{ps}$ increases with

the covariate imbalance. Thus, $\text{var}(\hat{\tau}_{ps} \mid N_{T1} = 40) > \text{var}(\hat{\tau}_{ps})$ and the observed test statistics are more spread out than the reference distribution they are being compared to.

The unconditional properties supported the notion that we should adjust for covariate imbalance either by modifying the test statistic or by conditioning. The conditional properties indicate that modifying the test statistic is inferior to conditioning because unconditional tests with modified test statistics can have the wrong conditional signficance level.

2.6 Product marketing example

Finally, we return to the product marketing example that initially motivated this exploration of conditional randomization tests. Remember that the experiment involved roughly 2000 experimental subjects and $K = 11$ treatment levels, which were the eleven versions of a particular product. Each subject randomly received by mail one of products. Each subject used the product and returned a survey regarding the product's performance. The outcome of interest was an ordinal variable with three levels, 1, 2, and 3 (with 3 being the best), and the goal was to identify which product version the subjects preferred. The survey also collected covariate information, including income and ethnicity and the experimenters were concerned about the effect of covariate imbalance on their conclusions. Critically, the covariate information was not collected until after the units were assigned to treatment levels and thus blocking and rerandomization were not possible.

More precisely, after removing observations with missing values, there were $N =$

2256 experimental units. The number of units assigned to each treatment level is given Table 2.4 (and the percentage below that).

Table 2.4: **Number of units assigned to each treatment level:** The number of units assigned to each treatment level was relatively equal.

	Treatment										
	1	2	3	4	5	6	7	8	9	10	11
# of Units	238	266	225	231	237	226	198	135	136	136	228
Percentage	10%	12%	10%	10%	11%	10%	9%	6%	6%	6%	10%

The analysis was broken into an omnibus test and a set of pairwise tests and all tests were carried out conditionally. In the omnibus test, we test the sharp null hypothesis that all K unit level potential outcomes are equal, $H_0 : Y_i(1) = \cdots = Y_i(11)$ for all $i = 1, \ldots, N$. If we reject the sharp null, we move on to the pairwise tests, where we compare all $\binom{11}{2} = 55$ pairs of treatments to determine which treatment the subjects preferred.

For this analysis, we consider the following eight covariates, all of which are categorical.

1. Order of detergent (3 levels)
2. Under stream (2 levels)
3. Care for dishes (5 levels)
4. Water hardness (5 levels)
5. Consumer segment (4 levels)

6. Household income (11 levels)

7. Age (6 levels)

8. Hispanic (2 levels)

We used these covariates to create clusters of the N observations. Because the covariates are categorical, we use the k-modes algorithm introduced by Huang (1997). This clustering method extends the k-means algorithm to handle categorical variables. The k-modes algorithm relies on a dissimarity measure, $d(\cdot, \cdot)$, which measures the dissimilarity between two observations. The dissimarity measure is the number of categorical variables which are different between the two obervations. So, if $X_i = (1, 2, 4, 2, 1, 10, 3, 1)$ and $X_j = (2, 1, 4, 2, 1, 10, 3, 1)$, then $d(X_i, X_j) = 2$. The smaller the dissimilarity measure the more similar the two observations. This is a simple dissimilarity measure in the sense that it gives equal weight to all covariates and completely ignores the ordinal structure of some of the categorical variables. For instance, an income value of 11 is much closer to an income value of 10 than to 1 but this aspect is ignored by this measure. Modifications to the dissimarily measure are left to future work. The mode of a set of observations, $\{X_1, \ldots, X_n\}$, is the vector Q that minimizes

$$\sum_{i=1}^{n} d(Q, X_i). \tag{2.43}$$

The k-modes algorithm follows the familar steps of the k-means algorithm. We start with k candidate modes. We then assign each observation to the closest mode according to the dissimilarity measure. We then re-calculate the modes of each cluster

and repeat these last two steps until convergence. We can determine an appropriate number of clusters, k, via an elbow plot, shown in Figure 2.5.

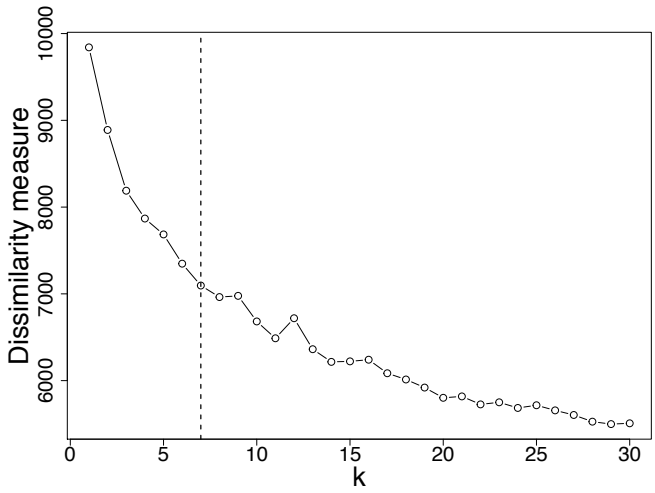


Figure 2.5: **Elbow plot:** Although somewhat arbitrary, we determined that the elbow is at $k = 7$, the vertical dashed line.

In this case, we decided to let $k = 7$. The contingency table, in Table 2.5, summarizes the number of units in each cluster assigned to each treatment level.

Table 2.5: **Clusters and treatment levels:** The rows are the seven clusters and the columns are the elevent treatment levels.

	Treatment											
	1	2	3	4	5	6	7	8	9	10	11	
1	82	93	63	88	83	84	71	39	56	46	78	783
2	35	28	29	26	28	25	21	13	12	16	27	260
3	44	37	41	47	34	37	30	32	22	23	44	391
4	21	29	28	18	22	20	10	17	12	13	21	211
5	14	26	20	20	18	18	14	8	9	11	22	180
6	16	17	22	13	24	17	24	11	10	11	11	176
7	26	36	22	19	28	25	28	15	15	16	25	255

The advantage of the clustering method is that we can replace the eight categorical covariates with one categorical covariate, the cluster indicator. As we have seen, it is particularly easy to sample treatment assignment when there is a single categorical covariate. We consider clustering a simple but useful first step in carrying out a conditional randomization test.

2.6.1 Omnibus Test

With the clusters in hand, we test the sharp null hypothesis that all K unit level potential outcomes are equal, $H_0 : Y_i(1) = \cdots = Y_i(11)$ for all $i = 1, \ldots, N$. We use the Kruskal-Wallis statistic as the test statistic (Kruskal and Wallis, 1952). The Kruskal-Wallis statistic is used in the Kruskal-Wallis test, a non-parametric test similar to one-way ANOVA. The statistic is similar to the F-statistic in that it is a ratio of sum of squares and that larger values of the statistic indicate that the treatment levels are different. The statistic is found by first computing the ranks of

the observed outcomes, $\mathbf{r}^{\text{obs}}$. The test statistic is then

$$(N-1)\frac{\sum_{k=1}^{K} N_k(\bar{r}_k^{\text{obs}} - \bar{r}^{\text{obs}})^2}{\sum_{i=1}^{N}(r_i^{\text{obs}} - \bar{r}^{\text{obs}})^2}, \tag{2.44}$$

where $\bar{r}_k^{\text{obs}}$ is mean rank in the kth treatment level and $\bar{r}^{\text{obs}}$ is the mean rank overall. Note that in this context, k indexes the treatment level and is not the number of clusters.

We carry out the conditional randomization test by conditioning on the number of units in each cluster assigned to each treatment level. Figure 2.6 reports the observed value of the Kruskal-Wallis test statistic (red line) and the reference distribution.

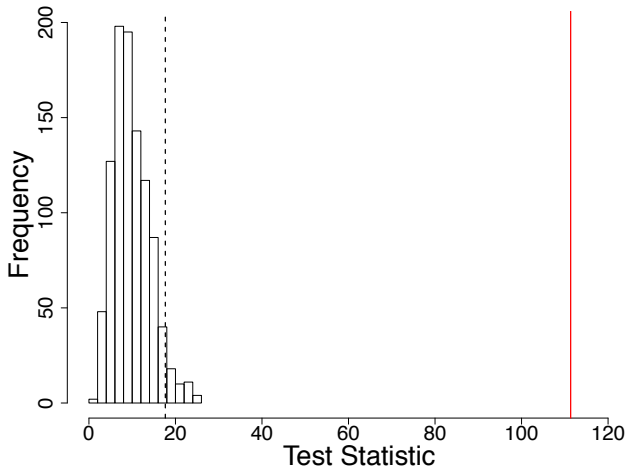


Figure 2.6: **Conditional randomization test using Kruskal-Wallis test statistic:** The vertical red line is the observed value of the test statistic. The histogram is the conditional distribution of the test statistic under the sharp null hypothesis.

The 0.95 quantile of the reference distribution is 18.05. The observed test statistic is 111.4 and the p-value is approximately 0. Thus, we strongly reject the sharp null

that $Y_i(1) = \cdots = Y_i(11)$ for all $i = 1, \ldots, N$.

2.6.2 Pairwise Tests

Having rejected the sharp null in the omnibus test, we next test for specific pairwise differences between the treatment levels. The mean outcome in each treatment level is shown in Table 2.6 and visualized in Figure 2.7 in decreasing order.

Table 2.6: **Mean outcome by treatment level:** Not adjusting for differnces in covariates, treatment 1 appears to be the most preferred treatment by the experimental subjects.

	Treatment										
	1	2	5	4	11	3	6	8	10	9	7
Mean	2.47	2.37	2.33	2.32	2.17	2.16	2.15	2.13	2.11	2.05	1.83

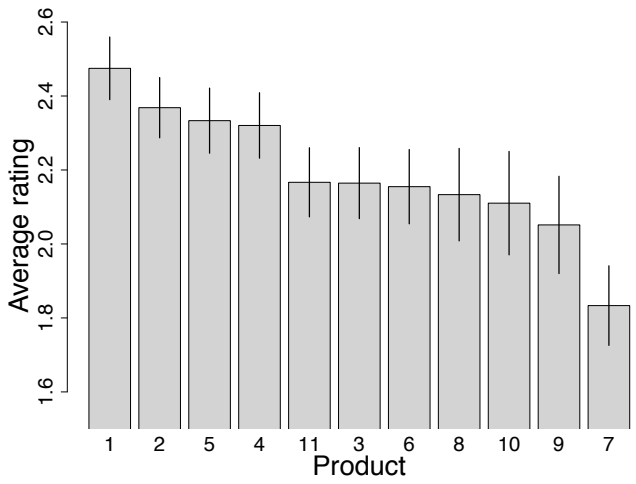


Figure 2.7: **Mean outcome by treatment level:** The vertical lines mark the 95% confidence intervals.

For the pairwise test between between treatment j and l, we test the sharp null hypothesis that $H_0 : Y_i(j) = Y_i(l)$ for all $i = 1, \ldots, N$. The two-sided p-values for the pairwise tests are shown in Table 2.7.

Table 2.7: ***p*-values for pairwise tests:** The 55 p-values show that for instance, that the difference between treatments 1 and 2 is just barely statistically significant.

	Treatment									
	2	5	4	11	3	6	8	10	9	7
1	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2		0.48	0.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5			0.81	0.02	0.01	0.02	0.01	0.02	0.00	0.00
4				0.02	0.01	0.03	0.02	0.03	0.00	0.00
11					0.92	0.91	0.63	0.68	0.18	0.00
3						0.99	0.61	0.50	0.19	0.00
6							0.58	0.66	0.26	0.00
8								0.99	0.61	0.00
10									0.60	0.00
9										0.01

The difference between treatments 1 and 2 is just barely statistically significant. We conclude that product version 1 is the most preferred product and that versions 1, 2, 5, and 4 are clearly preferred to the seven other products.

2.7 Conclusion

In this chapter, we considered conditional randomization tests as a form of covariate adjustment for randomized experiments. Conditional randomization tests have received relatively little attention in the statistics literature and we built upon Rosenbaum (2002) and Zheng and Zelen (2008) by introducing original notation to prove that the conditional randomization test has the correct unconditional significance level and to describe covariate balance more formally. Our simulation results verify that conditional randomization tests behave like more traditional forms of covariate

adjustment but have the added benefit of having the correct conditional significance level. While sampling treatment assingments is currently only feasible for the certain covariate balance functions, there are promising related methods and we feel that this challenge can be overcome.

Chapter 3

Bayesian optimal design of fixed knockout tournament brackets

3.1 Introduction

Conceptually, tournaments are a type of experimental design (Kendall, 1955). They are sequences of games betwen players designed to achieve some objective. The games are typically between two players and the objectives include, but are not limited to, identifying the best player. We will focus exclusively on knockout tournaments, which are among the most commonly used tournament structures. Examples include the tennis Grand Slam singles championships, the National Basketball Association (NBA) playoffs, and the knockout stage of the FIFA World Cup Finals. Perhaps, the most well-known knockout tournament in the United States is the the annual NCAA basketball tournament. The 2013 tournament bracket is shown in Figure 3.1.

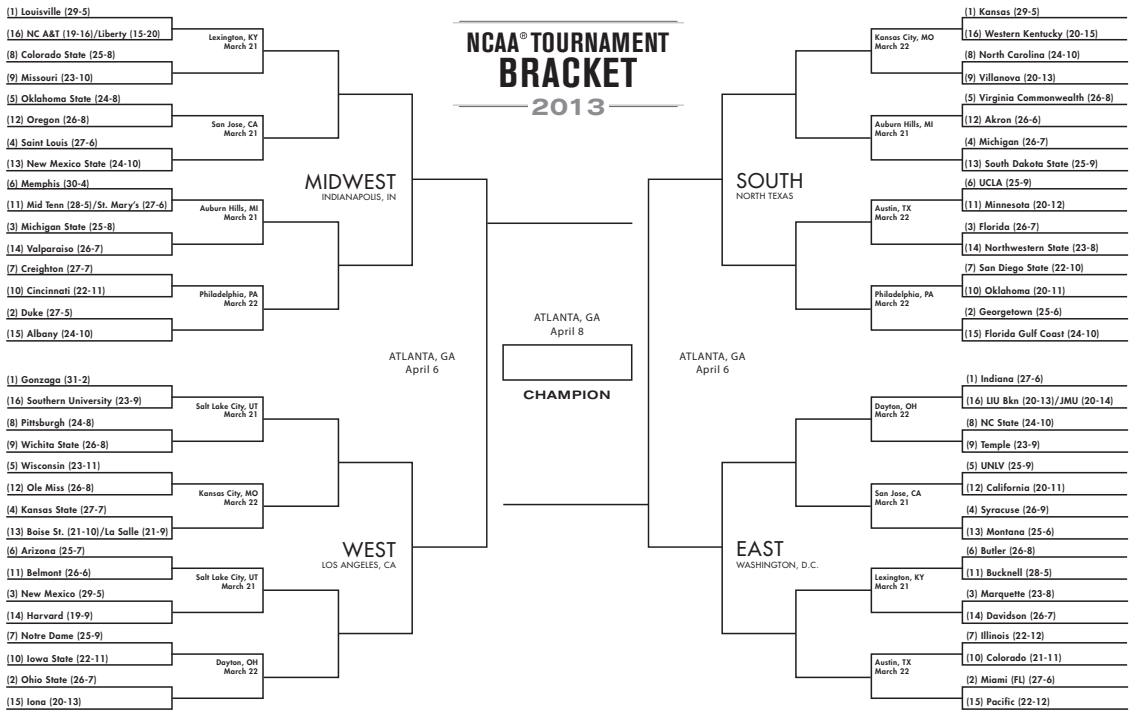


Figure 3.1: **2013 NCAA tournament bracket:** Note the tree structure of the bracket. For instance in the upper right corner of the bracket. The winners of the Kansas-Western Kentucky and North Carolina-Villanova games will meet in the next round regardless of the outcomes of the other games.

The purpose of tournament design is to choose the optimal tournament structure for a specific objective and we present a methodology for finding the optimal knockout tournament for a number of objectives (David, 1988).

Let N be the number of players in the tournament. Our first major assumption is to represent the strength of player i with a single number, θ_i. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ be the vector of player strengths for all N players. The larger θ_i, the stronger the ith player and the probability player i defeats player j is a function of $\theta_i - \theta_j$. Here,

we assume the function is $\Phi(\theta_i - \theta_j)$. Most tournament design methods treat $\boldsymbol{\theta}$ as a known quantity and find the tournament that achieves some objective. This is reasonable but some objectives are not well justified in this case. For instance, a popular tournament objective and one we will explore extensively is identifying the best player. However, if $\boldsymbol{\theta}$ is known, then we know before the tournament begins that the best player is player i^*, where $i^* = \text{argmax}_i \theta_i$. In that case, the tournament is unnecessary.

Following Glickman (2008), we treat the player strengths as uncertain. We represent this uncertainty by assuming a prior probability distribution on $\boldsymbol{\theta}$, $\text{p}(\boldsymbol{\theta})$. In this case, we do not know which player is the best and rather than being unncessary, the tournament is a tool to help us identify the best player. This is also a more realistic assumption. Even after a long season, player strengths are never known with certainty. There is always some level of statistical noise. Additionally, player strengths change over time and it is worth remembering that the notion of player strength is a simplification of a more complicated phenomenon. In this chapter, we introduce new design methods for knockout tournaments when player strengths are uncertain.

We focus on the cases where $N = 4$, 8, or 16, but the ideas extend to tournaments with larger numbers of players. In Section 3.2, we review the knockout tournament structure and paired comparison models. In Section 3.3, we introduce our Bayesian optimal design approach for finding the optimal bracket. In Section 3.4, we apply our approach to find the bracket that maximizes the probability that the best player wins the tournament and in Section 3.5, we consider other utility functions. In Section 3.6, we conclude.

3.2 Tournament background

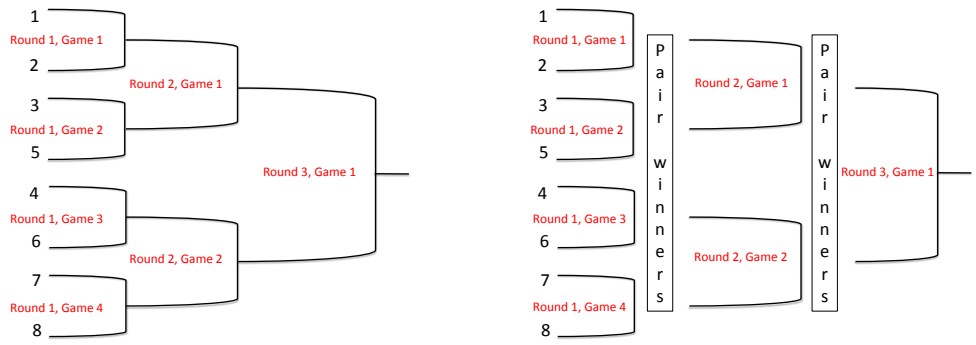
3.2.1 Knockout tournaments

In a knockout tournament, the number of players is a power of 2, $N = 2^R$, where R is the number of rounds. In the NCAA tournament in Figure 3.1, there are $R = 6$ rounds and $N = 2^6 = 64$ teams. In the first round, the N players are matched up in $N/2$ games. The $N/2$ first round winners advance to the second round and the $N/2$ losers are eliminated. Note that ties are not allowed. In the next round, the remaining $N/2$ players are matched up in $N/4$ games and this process continues recursively for R rounds. After the Rth round, only a single player remains, the tournament winner. The tournament winner will have won R games and lost 0. Since every player but the tournament winner loses exactly one game, there are $N - 1$ games in total.

We often refer to knockout tournaments as brackets and there are two types of brackets, fixed and adpative. In a fixed bracket, we know the complete structure of the tournament before it begins. For instance, in the fixed bracket shown in Figure 3.2(a), we know that the winners of the first and second games in round 1 will be matched up together in round 2 regardless of who the remaining players are. A fixed bracket has the traditional tree-structure shown in Figure 3.1 and is defined by the vector of first round matchups. We represent the fixed bracket in Figure 3.2(a) as $((1, 2), (3, 5), (4, 6), (7, 8))$. In the adaptive bracket shown in Figure 3.2(b), whether the winners of the first and second games in round 1 are matched up in round 2 does depend on who the remaining players are. The NBA and NCAA basketball tournaments use a fixed bracket whereas the NFL and NHL playoff tournaments use

an adaptive bracket in which teams are "reseeded" after the first round. Reseeding typically matches up the best remaining team with the worst remaining team, where best and worst teams could be determined on the basis of the regular season standings. Glickman (2008) studied the adaptive bracket and his objective was to match up players to maximize the probability the best player advances to the next round.

While we focus exclusively on the the fixed bracket, the adaptive bracket offers two potential advantages. The first is greater flexibility. For instance, if the goal is to maximize the probability that the best player wins the tournament, it makes sense to match the best remaining player with the worst remaining player after every round. The other advantage of the adaptive bracket is that after each round, we have the option of using the earlier games to update our prior about each player's strength. A weak player that has defeated several stronger opponents is likely stronger than we initially thought. The adaptive bracket allows the matchups to be adjusted accordingly.



(a) Fixed bracket

(b) Adaptive bracket

Figure 3.2: **Fixed vs adaptive bracket:** In the fixed bracket, the winner traditionally advances to play the winner of the neighboring game. For instance, the winner of round 1, game 1 advances to play the winner of round 1, game 2. In the adaptive bracket, the winners advance and the tournament designer matches up the winners at the beginning of each round. For instance, depending on who wins in the first round, the winner of round 1, game 1 could play the winner of round 1, game 4.

The NCAA tournament bracket in Figure 3.1, consists of four sub-brackets (this term will be defined later), labelled Midwest, South, West, and East, of 16 teams each. Before the tournament, the 16 teams in each sub-bracket are ranked by relative strength from 1 to 16 by a committee of experts. Team 1 is the best team in the sub-bracket and team 16 is the worst. Those 16 teams are then paired together in a bracket that follows what is called the *standard seeding*. The resulting bracket for the 16 teams is $((1,16),(8,9),(4,13),(5,12),(2,15),(7,10),(3,14),(6,11))$. Note that team i plays team $(16-i+1)$ in the first round and if the better team wins each first round game, team j will plays team $(8-j+1)$ in the second round and so on.

If the bracket follows the standard seeding, then if the better team wins each first round game game, the best $N/2$ teams will advance to the second round. Conditional on the best $N/2$ teams advancing to the second round, if the better team wins each second round game, the best $N/4$ teams will advance to third round and so on. The standard seeding increases the chance that the best teams will meet later in the tournament, which typically adds greater suspense to the final rounds. While the standard seeding is pervasive it does "not overtly adhere to any clear statistical principle" (Glickman, 2008).

3.2.2 Paired comparison models

Paired comparison models were developed to infer the preference ordering of N objects from comparisons of two objects at a time. The results of the comparison between objects i and j are that either i was preferred to j or j was preferred to i. It is also possible that the two objects were preferred equally, but for simplicity, we focus on games where results are decisive. The results are recorded and the data analyzed by fitting a paired comparison model. A preference ordering of the objects, along with measures of uncertainty, are the output. Thurstone (1927) introduced the first paired comparison model and major contributions include Mosteller (1951), Bradley and Terry (1952), and Luce (1959).

Paired comparison models can naturally be applied to sports results where a comparison between objects i and j corresponds to a game between players i and j and the result of the comparison is the result of the game. The preference ordering corresponds to a ranking of the players by strength. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ be the

strength parameters for the N players. Then, let $Y_{ij} = 1$ if player i defeats player j and $Y_{ij} = 0$ if player j defeats player i. Also, let $\boldsymbol{P} = [P_{ij}]$ be the matrix of the pairwise probabilities of winning, where $P_{ij} = \Pr(Y_{ij} = 1)$ represents the probability team i defeats team j. In the Thurstone-Mosteller model, we assume that

$$P_{ij} = \Pr(Y_{ij} = 1 \,|\, \boldsymbol{\theta}) = \Phi(\theta_i - \theta_j) \tag{3.1}$$

where $\Phi(\cdot)$ is the standard normal distribution function. In what follows, we assume the Thurstone-Mosteller model to be consistent with Glickman (2008). We also assume that the games are independent, so that

$$P(Y_{ij} = 1,\, Y_{ik} = 1 \,|\, \boldsymbol{\theta}) = \Phi(\theta_i - \theta_j)\, \Phi(\theta_i - \theta_k). \tag{3.2}$$

The only difference between the Thurstone-Mosteller and Bradley-Terry models is that the Bradley-Terry model assumes

$$P_{ij} = \Pr(Y_{ij} = 1 \,|\, \boldsymbol{\theta}) = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}}. \tag{3.3}$$

Inference centers on estimating $\boldsymbol{\theta}$. However, note that in both models $\boldsymbol{\theta}$ is non-identified since, for any value of $\boldsymbol{\theta}$, $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta} + c)$, where $L(\cdot)$ is the likelihood function for the model. This is easily solved by imposing a constraint on $\boldsymbol{\theta}$, such as $\sum_{i=1}^{N} \theta_i = 0$ or $\theta_N = 0$. Additionally, non-identifiability is not an issue in a Bayesian model where a proper prior is assumed on $\boldsymbol{\theta}$.

3.3 Optimal bracket methodology

Before describing our approach for finding the optimal fixed bracket, we first introduce the relevant notation. We then provide a general formula for the expected utility which we are interested in maximizing. We cover different approaches to calculate the expected utility and review simulated annealing, as a way to search through the many possible tournament brackets.

3.3.1 Notation

Let Θ be the set of all possible vectors of $\boldsymbol{\theta}$. Prior to the start of the tournament, we assume that knowledge about player strengths can be represented as a multivariate normal distribution,

$$\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3.4}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ is the vector of means, and $\boldsymbol{\Sigma}$ is the positive-definite covariance matrix with diagonal elements σ_i^2 and off-diagonal elements σ_{ij}. From the data generating process perspecitve, we think of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ as being drawn at the start of the tournament. We assume the player strengths are then constant throughout the tournament.

We represent a knockout tournament bracket, $\boldsymbol{b} = (b_1, b_2, \ldots, b_{\frac{N}{2}})$, as a vector of first round games. Here, $b_i = (b_{i1}, b_{i2})$ is the ith first round game and b_{i1} and b_{i2} are the indices of the two players in the ith game. Using this notation, we can write $\boldsymbol{b}$ as follows.

$$\boldsymbol{b} = ((b_{11}, b_{12}), \ldots, (b_{\frac{N}{2}1}, b_{\frac{N}{2}2})) \tag{3.5}$$

As another example, in the 8-player bracket in Figure 3.3,

$$\boldsymbol{b} = ((1,2),(3,5),(4,6),(7,8)). \tag{3.6}$$

However, note that this notation is not unique. For instance, when $N = 4$, the bracket $((1,2),(3,4))$ is the same as $((3,4),(1,2))$. We propose a canonical form for each bracket to avoid duplicate representations. We first define a sub-bracket. A sub-bracket is a vector of g "neighboring" first-round games, where g is a power of 2 such that $g = 1, 2, 4, 8, \ldots N/2$. There are thus $2g$ players in the sub-bracket. What makes the collection of games a sub-bracket is that after round $(\log_2 g + 1)$, exactly one of the players remains in the tournament. For example, in bracket $\boldsymbol{b}$ in Equation (3.6), ((1,2),(3,5)) is a sub-bracket, but ((3,5),(4,6)) is not. If $g = 1$, one of the two players in the game will be in the tournament after round $\log_2 1 + 1 = 1$. If $g = 8$, one of the 16 players will remain after round $\log_2 g + 1 = 4$. For instance, in Figure 3.3, (b_3, b_4) forms a two-game sub-bracket, and after round 2, only player 6 remains.

Any g-game sub-bracket where $g \geq 2$ can be divided into two $g/2$-game sub-brackets. In order to avoid counting the same bracket multiple times, for all sub-brackets with $g \geq 2$ games, the $g/2$-game sub-bracket with the player with the lowest index must be listed first. For a sub-bracket consisting of a single game, the player with the lowest index must be listed first (i.e. $(1,2)$ and not $(2,1)$) Brackets that statisfy this criterion are said to be in canonical form.

For instance, the following bracket $\boldsymbol{b}$,

$$\boldsymbol{b} = ((1,2),(3,5),(7,8),(4,6)). \tag{3.7}$$

is equivalent to the bracket in Figure 3.3 but $\boldsymbol{b}$ is not written in canonical form. In the 2-game sub-bracket, $((7,8),(4,6))$, the one-game sub-bracket with the player with the lowest index, $(4,6)$, is listed second rather than first. The canonical form is $((1,2),(3,5),(4,6),(7,8))$. Let $\mathscr{B}$ be the set of all brackets in canonical form and let $k = |\mathscr{B}|$ be the number of brackets. Searls (1963) showed that for an N player bracket, $k = |\mathscr{B}| = \frac{N!}{2^{N-1}}$.

Let W_i be the number of games player i wins in the tournament. For instance, $W_i = 0$ implies player i loses in the first round and $W_i = R$ implies player i wins the tournament. Note that W_i is a random variable. Let $\boldsymbol{W} = (W_1, \ldots, W_N)$ be the vector of N random variables. We call $\boldsymbol{W}$ the *win vector*. For the example in Figure 3.3, $\boldsymbol{W} = (1,0,0,0,3,2,1,0)$ and since player 5 wins the tournament, $W_5 = 3$. The win vector, together with $\boldsymbol{b}$, completely specifies the outcome of every game in the tournament.

Let $\Pr(\boldsymbol{W} = \boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b})$ be the probability $\boldsymbol{W} = \boldsymbol{w}$ for a given value of $\boldsymbol{\theta}$ and bracket $\boldsymbol{b}$. Because the games are assumed independent, $\Pr(\boldsymbol{W} = \boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b})$ is the product of $N-1$ probabilities, one for each game. For the example in Figure 3.3,

$$\begin{aligned}\Pr(\boldsymbol{W} = \boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b}) = {} & \Phi(\theta_1 - \theta_2)\, \Phi(\theta_3 - \theta_5)\, \Phi(\theta_4 - \theta_6)\, \Phi(\theta_7 - \theta_8) \\ & \Phi(\theta_1 - \theta_5)\, \Phi(\theta_6 - \theta_7)\, \Phi(\theta_5 - \theta_6).\end{aligned} \tag{3.8}$$

Let $\mathscr{W}_{\boldsymbol{b}}$ be the set of potentially observable win vectors for bracket $\boldsymbol{b}$. Note that $\mathscr{W}_{\boldsymbol{b}}$ depends on $\boldsymbol{b}$ since, for example, if players i and j are paired together in the first round, W_i and W_j cannot both be 0.

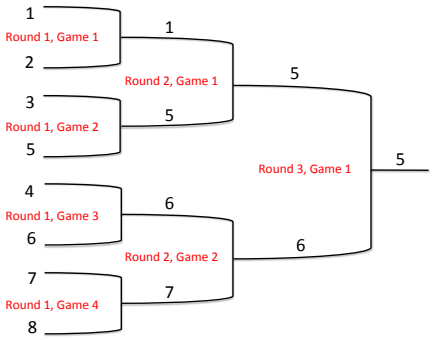


Figure 3.3: **Example bracket where $N = 8$:** For this bracket $\boldsymbol{b} = ((1,2),(3,5),(4,6),(7,8))$. And, for this example, $\boldsymbol{W} = (1,0,0,0,3,2,1,0)$.

Because there are $N-1$ games, the number of potentially observable win vectors corresponding to bracket $\boldsymbol{b}$ is $m = |\mathscr{W}_{\boldsymbol{b}}| = 2^{N-1}$. Remember that given $\boldsymbol{b}$, there is a one-to-one mapping between $\mathscr{W}_{\boldsymbol{b}}$ and the set of game outcomes in a knockout tournament.

The values of R, k, and m are shown for a few values of $N = 2^R$ in Table 3.1. Both k and m increase dramatically as N increases and this presents challenges in identifying the optimal bracket.

Table 3.1: **Number of rounds (R), brackets (k), and win vectors (m):** k and m increase dramatically as number of players (N) increases.

N	R	$k = \lvert\mathscr{B}\rvert$	$m = \lvert\mathscr{W}_b\rvert$
4	2	3	8
8	3	315	128
16	4	6.4×10^8	32,768
32	5	1.2×10^{26}	2.1×10^9
64	6	1.4×10^{70}	9.2×10^{18}

3.3.2 Expected utility

Following Lindley (1972), we apply the Bayesian decision-theoretic framework for optimal designs by specifying a utility function, $u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta})$, that is averaged over the data space ($\mathscr{W}_{\boldsymbol{b}}$) and the parameter space (Θ) for each bracket $\boldsymbol{b}$. The result is the expected utility, $U(\boldsymbol{b})$. We then maximize the expected utility over all possible brackets, $\boldsymbol{b} \in \mathscr{B}$. $U(\boldsymbol{b})$ can be written in any of the following equivalent ways.

$$
\begin{aligned}
U(\boldsymbol{b}) &= \sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} \int_{\Theta} u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}) \, \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{w}, \boldsymbol{b}) \, d\boldsymbol{\theta} \, \mathrm{p}(\boldsymbol{w}) \\
U(\boldsymbol{b}) &= \sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} \int_{\Theta} u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}) \, \mathrm{p}(\boldsymbol{w} \mid \boldsymbol{\theta}, \boldsymbol{b}) \, \mathrm{p}(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \\
U(\boldsymbol{b}) &= \int_{\Theta} \sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}) \, \mathrm{p}(\boldsymbol{w} \mid \boldsymbol{\theta}, \boldsymbol{b}) \, \mathrm{p}(\boldsymbol{\theta}) \, d\boldsymbol{\theta}.
\end{aligned} \tag{3.9}
$$

We introduce a variety of utility functions in Section 3.5, but we focus primarily on the utility function for the best player winning the tournament. The logic of

having the best player win the tournament goes back to the idea that the tournament is a tool and we are using it to help us identify the best player. To define the utility function, we first re-order the values of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ in decreasing order, $\theta_{(1)} > \theta_{(2)} > \cdots > \theta_{(N)}$, such that player (1) is the best, player (2) is the second best, and player (N) is the worst. While w_i is the number of wins for player i, let $w_{(1)}(\boldsymbol{\theta})$ be the number of wins for the best player, where the dependence on $\boldsymbol{\theta}$ is emphasized for readability. Then, the utility function for the best player winning the tournament is

$$u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}) = I(w_{(1)}(\boldsymbol{\theta}) = R). \tag{3.10}$$

In this case,

$$\begin{aligned} \sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b}) &= \sum_{\boldsymbol{w} \in \mathscr{W}_b} I(w_{(1)}(\boldsymbol{\theta}) = R)\, \mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b}) \\ &= \mathrm{p}(w_{(1)}(\boldsymbol{\theta}) = R \,|\, \boldsymbol{\theta}, \boldsymbol{b}) \end{aligned} \tag{3.11}$$

is the probability the best player wins the tournament for a given $\boldsymbol{\theta}$ and

$$U(\boldsymbol{b}) = \int_{\Theta} \mathrm{p}(w_{(1)}(\boldsymbol{\theta}) = R \,|\, \boldsymbol{\theta}, b)\, \mathrm{p}(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{3.12}$$

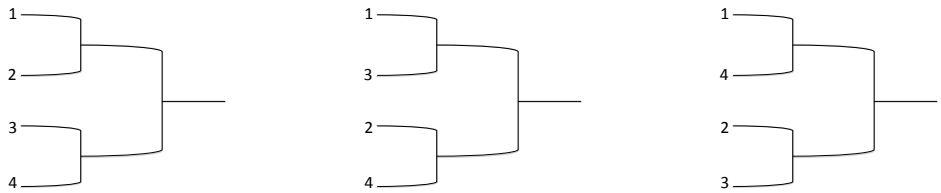
is the the probability the best player wins the tournament integrating over $\mathrm{p}(\boldsymbol{\theta})$. Finding the $\boldsymbol{b}$ that maximizes $U(\boldsymbol{b})$ is the same a maximizing the probability that the best player wins.

The overall computation time needed to find the optimal bracket is made up of two components. The first is the time needed to calculate $U(\boldsymbol{b})$. The second is the number of brackets for which we need to calculate $U(\boldsymbol{b})$. We next discuss both components.

3.3.3 Direct calculation of $U(\boldsymbol{b})$

We show how the expected utility, $U(\boldsymbol{b})$, can be calculated directly when $N = 4$. The same steps apply for larger N, but as we shall see, the computational time become prohibitively large when $N \geq 16$.

In Figure 3.4, we show the $k = 3$ brackets when $N = 4$.



(a) $((1,2),(3,4))$ (b) $((1,3),(2,4))$ (c) $((1,4),(2,3))$

Figure 3.4: **Brackets when $N = 4$:** There are only three brackets when $N = 4$.

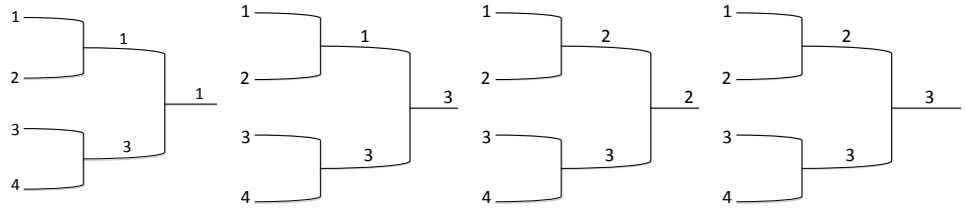
Following the formulation in Glickman (2008), we represent $U(\boldsymbol{b})$ as

$$U(\boldsymbol{b}) = \sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} \int_{\Theta} I(w_{(1)}(\boldsymbol{\theta}) = R)\, \mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b})\, \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{3.13}$$

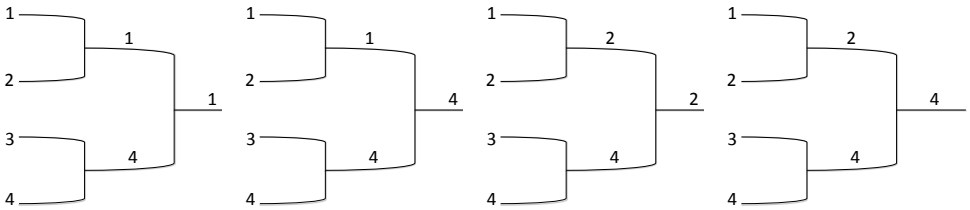
For each $\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}$, we calculate $\int_{\Theta} I(w_{(1)}(\boldsymbol{\theta}) = R)\, \mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b})\, \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ and sum over $\mathscr{W}_{\boldsymbol{b}}$.

The $m = 8$ win vectors for $\boldsymbol{b} = ((1,2),(3,4))$ are shown in Figure 3.5 and as

an example, we find $\int_\Theta I(w_{(1)}(\boldsymbol{\theta}) = R)\, \mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b})\, \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ for $\boldsymbol{w} = (2, 0, 1, 0)$. The integrals for the other $\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}$ follow similarly.



(a) $w = (2, 0, 1, 0)$ (b) $w = (1, 0, 2, 0)$ (c) $w = (0, 2, 1, 0)$ (d) $w = (0, 1, 2, 0)$



(e) $w = (2, 0, 0, 1)$ (f) $w = (1, 0, 0, 2)$ (g) $w = (0, 2, 0, 1)$ (h) $w = (0, 1, 0, 2)$

Figure 3.5: **Win vectors for $\boldsymbol{b} = ((1, 2), (3, 4))$**

Since $w_1 = 2$ when $\boldsymbol{w} = (2, 0, 1, 0)$, player 1 wins the tournament and thus, in order that $I(w_{(1)}(\boldsymbol{\theta}) = R) = 1$, player 1 must be the best player. This is equivalent to $I(\theta_1 > \theta_2,\, \theta_1 > \theta_3,\, \theta_1 > \theta_4) = 1$ and note that

$$\begin{aligned} I(\theta_1 > \theta_2, \theta_1 > \theta_3, \theta_1 > \theta_4) &= \Pr(\theta_1 > \theta_2,\, \theta_1 > \theta_3,\, \theta_1 > \theta_4 \,|\, \boldsymbol{\theta}) \\ &= \Pr(\theta_1 - \theta_2 > 0,\, \theta_1 - \theta_3 > 0,\, \theta_1 - \theta_4 > 0 \,|\, \boldsymbol{\theta}). \end{aligned} \quad (3.14)$$

Also, remember that since we are assuming a Thurstone-Mosteller model,

$$\begin{aligned}
\mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b}) &= \Phi(\theta_1 - \theta_2)\, \Phi(\theta_3 - \theta_4)\, \Phi(\theta_1 - \theta_3) \\
&= \Pr(z_1 < \theta_1 - \theta_2,\, z_2 < \theta_3 - \theta_4,\, z_3 < \theta_1 - \theta_3 \,|\, \boldsymbol{\theta}) \\
&= \Pr(\theta_1 - \theta_2 - z_1 > 0,\, \theta_3 - \theta_4 - z_2 > 0,\, \theta_1 - \theta_3 - z_3 > 0 \,|\, \boldsymbol{\theta}) \qquad (3.15)
\end{aligned}$$

where z_1, z_2, and z_3 are independent $N(0,1)$ random variables.

We can then show that $\int_\Theta I(w_{(1)}(\boldsymbol{\theta}) = R)\, \mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b})\, \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ equals the following.

$$\begin{aligned}
&\int_\Theta I(w_{(1)}(\boldsymbol{\theta}) = R)\, \mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b})\, \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \\
&\quad \int_\Theta \Pr(\theta_1 - \theta_2 > 0, \theta_1 - \theta_3 > 0, \theta_1 - \theta_4 > 0 \,|\, \boldsymbol{\theta}) \mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b})\, \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&\quad = \int_\Theta \Pr(\theta_1 - \theta_2 > 0, \theta_1 - \theta_3 > 0, \theta_1 - \theta_4 > 0 \,|\, \boldsymbol{\theta}) \\
&\quad \Pr(\theta_1 - \theta_2 - z_1 > 0,\, \theta_3 - \theta_4 - z_2 > 0,\, \theta_1 - \theta_3 - z_3 > 0 \,|\, \boldsymbol{\theta}) \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&\quad = \int_\Theta \Pr(\theta_1 - \theta_2 > 0, \theta_1 - \theta_3 > 0, \theta_1 - \theta_4 > 0, \\
&\quad \theta_1 - \theta_2 - z_1 > 0,\, \theta_3 - \theta_4 - z_2 > 0,\, \theta_1 - \theta_3 - z_3 > 0 \,|\, \boldsymbol{\theta}) \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&\quad = \Pr(\theta_1 - \theta_2 > 0,\, \theta_1 - \theta_3 > 0,\, \theta_1 - \theta_4 > 0, \\
&\quad \theta_1 - \theta_2 - z_1 > 0,\, \theta_3 - \theta_4 - z_2 > 0,\, \theta_1 - \theta_3 - z_3 > 0) \qquad (3.16)
\end{aligned}$$

The third equality holds because conditional on $\boldsymbol{\theta}$, functions of $\boldsymbol{\theta}$ are independent. We can calculate the last expression exactly by letting $\boldsymbol{\gamma} = (\boldsymbol{\theta}, z_1, z_2, z_3)$, which is a multivariate normal vector. Then, each component of the last expression is a linear

transformation of $\boldsymbol{\gamma}$. In fact, let $\tilde{\boldsymbol{\gamma}}$ be

$$\tilde{\boldsymbol{\gamma}} = (\theta_1 - \theta_2,\, \theta_1 - \theta_3,\, \theta_1 - \theta_4,\, \theta_1 - \theta_2 - z_1,\, \theta_3 - \theta_4 - z_2,\, \theta_1 - \theta_3 - z_3), \tag{3.17}$$

where $\tilde{\boldsymbol{\gamma}}$ is also multivariate normal. Then,

$$\int_{\Theta} I(w_{(1)}(\boldsymbol{\theta})) = R)\, \mathrm{p}(\boldsymbol{w} \mid \boldsymbol{\theta}, \boldsymbol{b})\, \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int_{\tilde{\boldsymbol{\gamma}}>0} \mathrm{p}(\tilde{\boldsymbol{\gamma}})\, d\tilde{\boldsymbol{\gamma}}. \tag{3.18}$$

The integral over the positive orthant can be calculated using the method described in Genz (1992). These integrals must be carried out for all m win vectors in $\mathscr{W}_{\boldsymbol{b}}$ and the sum of the probabilities is $U(\boldsymbol{b})$. On a 3.2GHz quad-core CPU, it takes 0.05 seconds to calculate $U(\boldsymbol{b})$ for one bracket when $N = 4$ and 6 seconds when $N = 8$. This implies that it takes 30 minutes, a significant but not unreasonable amount of time, to find the optimal bracket when $N = 8$. However, when $N = 16$, it takes approximately 4.5 hours to calculate $U(\boldsymbol{b})$ for a single bracket. Since there are $k = 6.4 \cdot 10^8$ brackets, it would be infeasible to calculate the expected utility for every bracket when $N = 16$ using the described approach.

3.3.4 Estimating $U(\boldsymbol{b})$

We next consider two approaches for estimating $U(\boldsymbol{b})$ when $N = 16$. The first approach uses a random sample from Θ and the second uses a random sample from $\mathscr{W}_{\boldsymbol{b}}$.

Sampling from Θ

We can estimate $U(\boldsymbol{b})$ using Equation (3.9) and Monte Carlo integration by sampling from Θ. We represent $U(\boldsymbol{b})$ as

$$U(\boldsymbol{b}) = \int_{\Theta} \sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}) \, \mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}, \boldsymbol{b}) \, \mathrm{p}(\boldsymbol{\theta}) \, d\boldsymbol{\theta}. \tag{3.19}$$

Since, $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can easily generate a random sample of L draws, $\boldsymbol{\theta}^1, \ldots \boldsymbol{\theta}^L$. For many utility functions, we can calculate $\sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}^\ell) \, \mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}^\ell, \boldsymbol{b})$ for draw ℓ by finding the probability each player advances to each round. We then estimate $U(\boldsymbol{b})$ with

$$\widehat{U(\boldsymbol{b})} = \frac{1}{L} \sum_{\ell=1}^{L} \Big(\sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}^\ell) \, \mathrm{p}(\boldsymbol{w} \,|\, \boldsymbol{\theta}^\ell, \boldsymbol{b}) \Big). \tag{3.20}$$

The larger the value of L, the more precise the estimate of $U(\boldsymbol{b})$ and as we shall see, precision is important because many of the brackets have similar expected utilities.

For a given $\boldsymbol{\theta}$, we calculate the probability each player advances to each round of the tournament by first finding the probability that each player advances to the second round. The probability each player advances to the second round is just the probability that the player wins their first round game. The probability that a player advances to the third round is the probability the player advances to the second round multiplied by the probability the player wins his second round game. To find the probability of winning the second round game, we multiply the player's probability of defeating each possible second round opponent by the probability the opponent advances to the second round and sum the products over the possible opponents. We

continue to apply this process to the later rounds. The calculation reduces to a series of R matrix multiplications, one for each round, that can be quickly carried out for all L draws simultaneously.

We also found that we can dramatically reduce the number of draws of $\boldsymbol{\theta}$ necessary to achieve a given precision by utilizing quasi-Monte Carlo integration (Morokoff and Caflisch, 1995). Quasi-Monte Carlo integration uses a low-discrepancy sequence, such as the Sobol sequence, to select the $\boldsymbol{\theta}$ draws. Using the low-discrepancy sequence means that the draws are deterministically chosen and thus, no longer randomly selected. The draws tend to be more evenly spread over the high probability regions and thus, the estimates are more precise. It is conceptually similar to space filling designs and quadrature methods. However, quadrature methods weight points differently and do not scale to larger dimensions. While the convergence rate for Monte Carlo integration is $O(\frac{1}{\sqrt{L}})$, the convergence rate for quasi-Monte Carlo is $O(\frac{1}{L})$.

Sampling from $\mathscr{W}_{\boldsymbol{b}}$

We can also estimate $U(\boldsymbol{b})$ by sampling from $\mathscr{W}_{\boldsymbol{b}}$. This is very similar to our direct calculation of $U(\boldsymbol{b})$ and we return to representing $U(\boldsymbol{b})$ as

$$U(\boldsymbol{b}) = \sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} \int_{\Theta} u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}) \, \mathrm{p}(\boldsymbol{w} \mid \boldsymbol{\theta}, \boldsymbol{b}) \, \mathrm{p}(\boldsymbol{\theta}) \, d\boldsymbol{\theta}. \tag{3.21}$$

In our direct calculation, we summed over all $\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}$ but now, we sum over a random sample of the win vectors, $\boldsymbol{w}^1, \ldots, \boldsymbol{w}^L$. Our estimate of $U(\boldsymbol{b})$ is then

$$\widehat{U(\boldsymbol{b})} = \frac{k}{L} \sum_{\ell=1}^{L} \Big(\int_{\Theta} u(\boldsymbol{b}, \boldsymbol{w}^{\ell}, \boldsymbol{\theta}) \, \mathrm{p}(\boldsymbol{w}^{\ell} \mid \boldsymbol{\theta}, \boldsymbol{b}) \, \mathrm{p}(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \Big). \tag{3.22}$$

We generate $\boldsymbol{w}^1, \ldots, \boldsymbol{w}^L$ by randomly simulating the result of each game in tournament with bracket $\boldsymbol{b}$ from a Bernoulli(0.5).

When the utility function is the best player winning the tournament, estimating $U(\boldsymbol{b})$ by sampling from $\mathscr{W}_{\boldsymbol{b}}$ tends to be slower than estimating $U(\boldsymbol{b})$ by sampling from Θ because the integral $\int_\Theta u(\boldsymbol{b}, \boldsymbol{w}^\ell, \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{w}^\ell \mid \boldsymbol{\theta}, \boldsymbol{b})\, \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ must be calculated L times while calculating $\sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}^\ell)\, \mathrm{p}(\boldsymbol{w} \mid \boldsymbol{\theta}^\ell, \boldsymbol{b})$ for all L draws requires only R matrix multiplications. As an example, we estimate the probability that the best player wins when $N = 16$ using both approaches. Let $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\begin{aligned} \boldsymbol{\mu} &= (0.75, 0.65, 0.55, 0.45, 0.35, 0.25, 0.15, 0.05, \\ &\quad -0.05, -0.15, -0.25, -0.35, -0.45, -0.55, -0.65, -0.75) \\ \boldsymbol{\Sigma} &= 0.1 \cdot I. \end{aligned} \tag{3.23}$$

Let the bracket follow the standard seeding based on the ordering of $\boldsymbol{\mu}$. The results are reported in Table 3.2. Sampling from Θ is over 100 times faster and leads to standard errors over 20 times smaller. Also, the standard error of $\widehat{U(\boldsymbol{b})}$, $\mathrm{se}(\widehat{U(\boldsymbol{b})})$, for sampling from $\mathscr{W}_{\boldsymbol{b}}$ is too large to be practically useful. In order to distinguish between brackets, the standard error needs to be at least as small as the reported standard error when sampling from Θ.

Table 3.2: **Comparing two estimates of $U(\boldsymbol{b})$:** We compare one estimate from each estimation method on a 3.2GHz quad-core CPU. Each method uses 3000 draws of either $\boldsymbol{\theta}$ or $\boldsymbol{w}$.

	$\widehat{U(\boldsymbol{b})}$	$\text{se}(\widehat{U(\boldsymbol{b})})$	Time (sec)
Sampling from Θ	0.3567	0.0020	0.07
Sampling from $\mathscr{W}_{\boldsymbol{b}}$	0.3972	0.0558	108

However, sampling from $\mathscr{W}_{\boldsymbol{b}}$ has the potential to be useful because some utility functions make sampling from Θ computationally infeasible. As we discuss later, some utility functions place severe restrictions on $\boldsymbol{\theta}$, such that most of the draws of $\boldsymbol{\theta}$ do not satisfy the restrictions. When sampling from Θ, we then end up estimating $U(\boldsymbol{b})$ to be 0 for many brackets, making it difficult to distinguish among brackets. However, these restrictions are easily handled when sampling from $\mathscr{W}_{\boldsymbol{b}}$ because we can incorporate the restrictions into $\tilde{\gamma}$ as we did previously.

3.3.5 Simulated annealing

When $N = 4$ or 8, the number of brackets is $k = 3$ or $k = 315$, respectively, and we can find the optimal bracket by calculating $U(\boldsymbol{b})$ for every bracket and selecting the maximum. However, when $N = 16$, the number of brackets is $k = 6.4 \cdot 10^8$ and it is computationally infeasible to calculate $U(\boldsymbol{b})$ for every bracket. In these cases, to find the optimal bracket, we use the Markov chain-based optimization method simulated annealing. Simulated annealing was introduced by Kirkpatrick et al. (1983) and is a method for finding a good approximation to the maximum of a function without

evaluating the function at every point.

In simulated annealing, we run a Metropolis algorithm on the bracket space, $\mathscr{B}$. Let $\pi(\boldsymbol{b}) \propto \exp(U(\boldsymbol{b})/T)$ be a probability distribution on $\mathscr{B}$ based on $U(\boldsymbol{b})$, where T is the temperature parameter and is initialized to a large value, $T = T_0$. We start with an initial bracket $\boldsymbol{b}^1$ and generate a proposal bracket, $\boldsymbol{b}^P$, from a pre-specified neighborhood around $\boldsymbol{b}^1$, which is accepted or rejected according to a Metropolis acceptance probability,

$$\mathrm{p_{acc}} = \min\Big(1, \frac{\pi(b^P)}{\pi(b^1)}\Big) = \min(1, e^{\frac{U(b^P)-U(b^1)}{T}}). \tag{3.24}$$

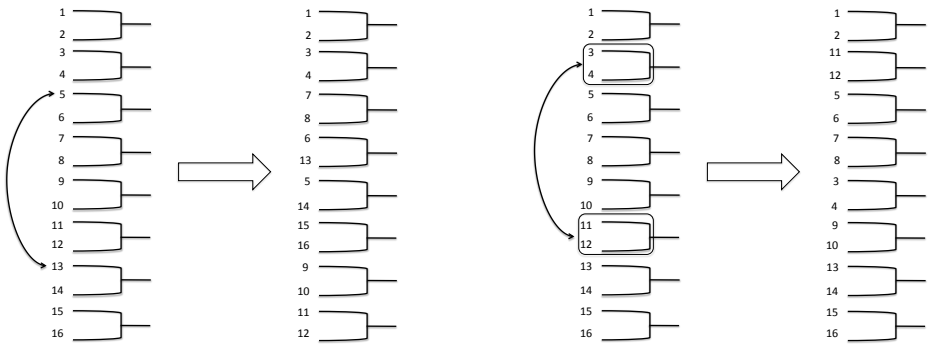
If the proposal is accepted, we let $\boldsymbol{b}^2 = \boldsymbol{b}^P$. If the proposal is rejected, we let $\boldsymbol{b}^2 = \boldsymbol{b}^1$. This process is repeated iteratively for $K_{\max}$ steps. The temperature T is lowered after each step according to an *annealing schedule*. The temperature can also be lowered after a fixed number of steps to give the the chain a chance to converge at a fixed T. As the temperature is lowered, the chain is less likely to jump to brackets with lower expected utility. For instance, if at step i, if $U(\boldsymbol{b}^P) < U(\boldsymbol{b}^{i-1})$, then $\mathrm{p_{acc}} = e^{\frac{U(\boldsymbol{b}^P)-U(\boldsymbol{b}^{i-1})}{T}} < 1$. As T decreases, so does $\mathrm{p_{acc}}$. The lowering of the temperature is called *slow cooling*. As T gets closer to 0, simulated annealing reduces to the greedy algorithm which only makes moves that increase the expected utility. Once the algorithm is complete, the bracket with the largest expected utility is selected. Note that $K_{\max}$, the intial temperature, T_0, and the annealing schedule should be tuned for each utility function and prior distribution on $\boldsymbol{\theta}$.

To generate a proposal bracket, $\boldsymbol{b}^P$, we consider a neighborhood of brackets that corresponds to randomly applying one of two types of swaps to the current bracket.

1. **Swap two players:** We select and swap two players that are *not* matched up in the first round. Players that are closer together, in terms of being in the same sub-brackets, are swapped with higher probability. For instance, each player has probability $1/N$ of being selected. Conditional on the first player being selected, we are more likely to select the second player from the same 2-game sub-bracket than the same 8-game sub-bracket. This ensures that proposals tend to be similar to the current bracket. Alternatively, we could swap players with similar strengths, but it is not obvious what measure of strength similarity would be appropriate. If the proposed bracket is not in canonical form, it is converted to canonical form.

2. **Swap two sub-brackets:** First, a sub-bracket size, the number of games in the sub-bracket, is selected. Let g be the selected sub-bracket size, where $1 \leq g \leq N/8$. Then, two sub-brackets of g-games are selected and swapped. Smaller values of g are selected with higher probability and thus, we are more likely to swap two two-game sub-brackets than two four-game sub-brackets. For instance, when $N = 16$, we are twice as likely to swap two one-game sub-brackets than two two-game sub-brackets. Note that the sub-brackets need to be non-adjacent, which means that they cannot be combined to form a sub-bracket. If the proposed bracket is not in canonical form, it is converted to canonical form.

Figure 3.6 illustrates the two types of swaps. Note that sub-bracket swaps are equivalent to a series of player swaps. However, we found it beneficial to include sub-bracket swaps because they help the chain jump out of local modes.



(a) Swap players 5 and 13 (b) Swap games $(3, 4)$ and $(11, 12)$

Figure 3.6: **Two types of swaps:** In (a), two players, 5 and 13, are swapped and the resulting bracket is displayed in canonical form. In (b), two games (i.e. two one-game sub-brackets), $(3, 4)$ and $(11, 12)$, are swapped and again the resulting bracket is displayed in canonical form.

To find the optimal bracket, we run the simulated annealing algorithm several times, each time from a different starting bracket. If we reach the same or nearly the same optimal bracket from different starting brackets, we are confident the algorithm is finding approximately optimal brackets. Natural starting brackets are those that are optimal given a likely orderings of the players. For instance, say a likely ordering of the player follows the player indices, $\theta_1 > \theta_2 > \ldots > \theta_{16}$. The optimal bracket given this ordering is shown in Figure 3.7 and would be a natural starting bracket.

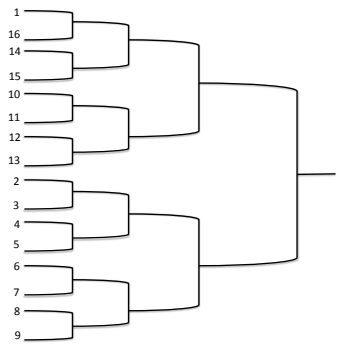


Figure 3.7: **Optimal bracket when $\theta_1 > \theta_2 > \ldots > \theta_{16}$:** This bracket could serve as a natural starting bracket for the simulated annealing algorithm.

In maximizing the probability of the best player winning when $N = 16$, we found the following parameter values worked well for a variety of priors. We let $K_{\max} = 2000$ and initially, let $T = 0.002$. At each iteration, we lowered T by a factor of $0.01^{1/K_{\max}}$. Also, in generating the proposal bracket, we swapped two players with probability 0.9, swapped two games with probability 0.07, and swapped two two-game sub-brackets with probability 0.03.

3.4 Maximizing the probability the best player wins the tournament

We find the brackets that maximize the probability of the best player winning the tournament for tournaments with different numbers of players and different prior distributions. We first apply our methodology to find the bracket $\boldsymbol{b}$ that maximizes $U(\boldsymbol{b})$ when $N = 4$. We compare the result to the optimal adaptive bracket in Glickman

(2008) and we briefly compare the $N = 4$ case to the $N = 8$ case. For the $N = 16$ case, we build on the simulation study in Glickman (2008) by comparing the optimal bracket to popular alternative brackets.

3.4.1 $N = 4$

We consider three prior distributions when $N = 4$. In all three, the prior mean is $\boldsymbol{\mu} = (0.09, 0.03, -0.03, -0.09)$ but the variances differ. We begin with the two simplest scenarios. The first is when the prior distribution is a point mass prior (i.e. the prior variance is 0) and the second is when the prior variances are equal. In these situations, there is a natural ordering of the players based on the prior means and the bracket that maximizes the probability the best player wins the tournament is the same in both cases. The player with the largest prior mean is matched up with the player with the smallest prior mean because the player with the largest prior mean is most likely the best player and we maximize that player's probability of winning the tournament by giving the player an easy first round game. This bracket also maximizes the probability that the best player advances to the next round, the utility function in Glickman (2008). However, when the prior variances are unequal, the optimal bracket changes and is no longer the same across different utility functions.

Point mass prior

We begin with a degenerate point mass prior on $\boldsymbol{\theta}$. In this case, $\boldsymbol{\theta}$ is known with certainty and $\boldsymbol{\theta} = (0.09, 0.03, -0.03, -0.09)$. Thus, player 1 is the best player and for each of the $k = 3$ brackets, we find the probability that player 1 wins the tournament.

The results are reported in Table 3.3.

Table 3.3: **Probability best player wins for point mass prior:** In this case, we know that player 1 is the best player and we report the probability that player 1 wins the tournament for the three different brackets.

Bracket	Prob. best player wins
$((1,2),(3,4))$	0.2929
$((1,3),(2,4))$	0.2987
$((1,4),(2,3))$	0.3059

In the case of a point mass prior, bracket $((1,4),(2,3))$ maximizes the probability that player 1 wins. Since we know player 1 is the best player it is intuitive that we should match up player 1 with the weakest possible opponent. Thus, in the first round, player 1 is matched up against player 4.

Equal variances

We next examine the equal variances example from Glickman (2008). Let

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} \sim N\left(\begin{pmatrix} 0.09 \\ 0.03 \\ -0.03 \\ -0.09 \end{pmatrix}, \begin{pmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & 0.3 \end{pmatrix} \right), \tag{3.25}$$

Note that θ_1, θ_2, θ_3, and θ_4 are independent, with different means but the the same variance. We visualize the prior distribution in Figure 3.8.

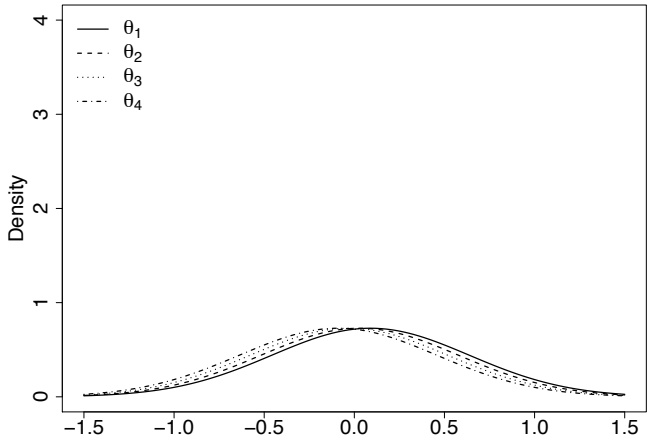


Figure 3.8: **Prior densities:** The distributions of θ_1, θ_2, θ_3, and θ_4 are largely similar because the prior standard deviation is significant compared to the differences in prior means.

The prior distributions are largely similar and the differences between consecutive prior means (0.06) are small compared to the prior standard deviations ($\sqrt{0.3} = 0.55$). When the players have equal variances, the prior means imply a natural ordering of the players. Player 1 has the largest probability of being the best and the lowest probability of being the worst. Player 2 has the second highest probability of being the best and the second lowest probability of being the worst and so on.

The probability that the best player wins for each bracket is shown in Table 3.4.

Table 3.4: **Probability best player wins for equal variance prior:** Bracket $((1,4),(2,3))$ maximizes the probability that the best player wins.

Bracket	Prob. best player wins
$((1,2),(3,4))$	0.5388
$((1,3),(2,4))$	0.5394
$((1,4),(2,3))$	0.5396

Because the prior distributions are similar, the probability that the best player wins is similar for the three brackets. Again, the optimal bracket is $((1,4),(2,3))$ and player 1, likely the best player, is matched up with player 4, likely the worst player. Also, bracket $((1,4),(2,3))$ remains the optimal bracket when the variances are increased or decreased, as long as they remain equal. As the prior variance decreases, the prior approaches the point mass prior discussed previously.

In Glickman (2008), the bracket $((1,4),(2,3))$ also maximized the probability that the best player wins their first round game. The same intuition applies. Player 1 is most likely the best and thus to maximize player 1's probability of making it to the second round, we should match up players 1 and 4.

Unequal variances

We next examine an example with unequal variances. The example is similar to Example 2 from Glickman (2008). Let

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} \sim N\left(\begin{pmatrix} 0.09 \\ 0.03 \\ -0.03 \\ -0.09 \end{pmatrix}, \begin{pmatrix} 1.0 & 0 & 0 & 0 \\ 0 & 1.0 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0.01 \end{pmatrix}\right), \tag{3.26}$$

Note that θ_1, θ_2, θ_3, and θ_4 are still independent and have the same means as before, but the variances are different. We visualize the prior distribution in Figure 3.9.

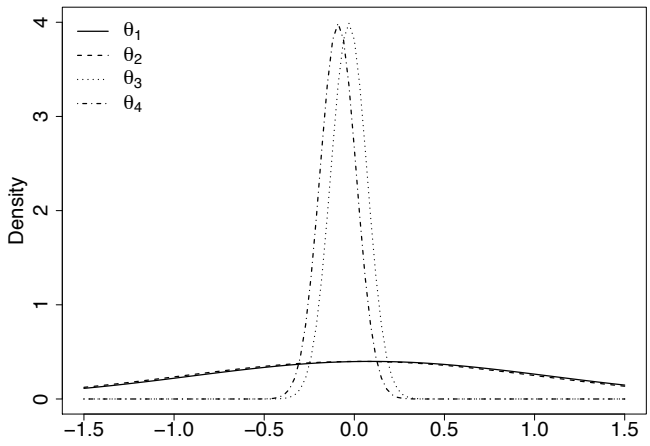


Figure 3.9: **Prior densities:** The prior distributions of θ_1 and θ_2 have much larger variances than the prior distributions of θ_3 and θ_4.

The prior means are evenly spaced but the prior variances for θ_1 and θ_2 are much larger than the prior variances for θ_3 and θ_4. We are much more uncertain about the strength of players 1 and 2 than of players 3 and 4. This could be because, for instance, players 1 and 2 have played fewer games we have less information about their performance. In Table 3.5, we report the probability each player is the best and the

probability each player is the worst. We also report the probability that each of the four players wins the tournament for each of the $k = 3$ brackets. These probabilities are found by slightly modifying the formulas discussed previously. Because the prior variances for θ_1 and θ_2 are so large, players 1 and 2 have the highest probabilities of being the best and also the highest probabilities of being the worst. The notion of the best player and the worst player is ambiguous when the prior variances are unequal and the intuition for why certain brackets are optimal is less obvious.

Table 3.5: **Bracket statistics for unequal variances example:** Note that players 1 and 2 have both the highest probabilities of being the best and worst players. Additionally, bracket $((1,2),(3,4))$ maximizes both the probabilities that players 1 and 2 win the tournament.

	Prob. player	Prob. player	Prob. player wins		
Player	is best	is worst	$((1,2),(3,4))$	$((1,3),(2,4))$	$((1,4),(2,3))$
1	0.401	0.323	0.344	0.322	0.325
2	0.371	0.350	0.321	0.305	0.301
3	0.153	0.108	0.180	0.196	0.198
4	0.075	0.219	0.156	0.178	0.175

Note that bracket $((1,2),(3,4))$ maximizes both the probabilities that players 1 and 2 win the tournament even though they play each other in the first round. In Table 3.6, we calculate the probability each player wins the tournament conditional on the player being the best.

Table 3.6: **Conditional probabilities of winning:** Conditional on being the best, bracket $((1,2),(3,4))$ maximizes the probabilities that players 1 and 2 win the tournament but minimizes the probabilities that players 3 and 4 win the tournament.

	Prob.	Conditional prob. player wins		
Player	player is best	$((1,2),(3,4))$	$((1,3),(2,4))$	$((1,4),(2,3))$
1	0.401	0.674	0.637	0.641
2	0.371	0.666	0.634	0.631
3	0.153	0.388	0.439	0.442
4	0.075	0.374	0.430	0.427

We can use Table 3.6 to find the probability of the best player winning the tournament for each bracket by multiplying the probability of each player being the best by the probability of winning the tournament conditional on being the best. For $\boldsymbol{b} = ((1,2),(3,4))$,

$$\begin{aligned} U(\boldsymbol{b}) &= 0.401 \cdot 0.674 + 0.371 \cdot 0.666 + 0.153 \cdot 0.388 + 0.075 \cdot 0.374 \\ &= 0.6049 \end{aligned} \tag{3.27}$$

In Table 3.7, we report the probability that the best player wins for each of the three brackets.

Table 3.7: **Probability best player wins for 3 brackets:** Bracket $((1,2),(3,4))$ maximizes the probability the best player wins the tournament.

Bracket	Prob. best player wins
$((1,2),(3,4))$	0.6049
$((1,3),(2,4))$	0.5906
$((1,4),(2,3))$	0.5909

While the probability the best player wins the tournament is very close for the three brackets, the bracket $((1,2),(3,4))$ maximizes the probability at 0.6049. Given that players 1 and 2 have the highest probabilities of being the best and that bracket $((1,2),(3,4))$ maximizes their probabilities of winning, this is not surprising. However, having the two players with the highest probabilities of being the best play each other in the first round is somewhat counterintuitive. We attempt to develop some intuition for the result by comparing it to the optimal adaptive bracket from Glickman (2008), where we maximize the probability the best player wins their first round game.

In Table 3.8, we calculate the probability each player wins their first round game conditional on the player being the best. We also calculate the probability each player wins their second round game conditional on winning their first round game and being the best player.

Table 3.8: **Conditional probabilities of winning first and second round games:** We report the probability of each player winning their first round game conditional on being the best and then the probability of winning their second round game conditional on winning their first round game and being the best player.

		Conditional prob. player wins					
	Prob.	$((1,2),(3,4))$		$((1,3),(2,4))$		$((1,4),(2,3))$	
Player	player is best	Rd. 1	Rd. 2	Rd. 1	Rd. 2	Rd. 1	Rd. 2
1	0.401	0.835	0.808	0.794	0.802	0.809	0.794
2	0.371	0.827	0.805	0.805	0.789	0.791	0.798
3	0.153	0.557	0.697	0.745	0.590	0.749	0.590
4	0.075	0.539	0.695	0.747	0.575	0.743	0.575

The bracket $((1,4),(2,3))$ maximizes the probability that the best player wins their first round game. This probability can be found by multiplying the probability each player is the best by the probability they win their first round game conditional on being the best. For $((1,4),(2,3))$, this is

$$0.401 \cdot 0.809 + 0.371 \cdot 0.791 + 0.153 \cdot 0.749 + 0.075 \cdot 0.743 = 0.7882.$$

However, as we have seen, bracket $((1,4),(2,3))$ does not maximize the probability that the best player wins the tournament. We can rewrite the probability the best player wins the tournament by multiplying the probability each player is the best by the probability they win their first round game conditional on being the best and by the probability they win their second round game conditional on winning their first round game and being the best. Again, for $((1,4),(2,3))$, this is

$$0.401 \cdot 0.809 \cdot 0.794 + 0.371 \cdot 0.791 \cdot 0.798 + 0.153 \cdot 0.749 \cdot 0.590 + 0.075 \cdot 0.743 \cdot 0.575 = 0.5909.$$

Multiplying by the second conditional probability is what makes the fixed bracket problem different from the adaptive bracket problem. Note that the second conditional probability involves a Bayesian update of the prior distribution, where the prior is updated based on the result of the previous game.

The reason bracket $((1,4),(2,3))$ does not maximize the probability of winning the tournament can be understood by again considering the conditional probabilities. Conditional on player i being the best, the relatively easier opponents are players 1 and 2 because there is a good chance that θ_1 and θ_2 are less than θ_i. This also holds if $i = 1$ or 2. Conversely, conditional on player i being the best, the relatively harder opponents are players 3 and 4. Thus, in $((1,4),(2,3))$, conditional on being the best, players 3 and 4 likely play an easy opponent in the first round and a hard opponent in the second round. Conditional on being the best, players 1 and 2 likely play a hard opponent in the first round and a hard opponent in the second round.

In bracket $((1,2),(3,4))$, again, players 3 and 4 likely play one easy and one hard opponent, although they play the hard opponent first. This means that conditional on being the best, players 3 and 4 have a lower chance of winning their first round game (0.557 versus 0.749 for player 3 and 0.539 versus 0.743 for player 4) but a higher probability of winning their second round game conditional on winning their first round game (0.697 versus 0.590 for player 3 and 0.695 versus 0.575 for player 4). Because the probabilities players 3 and 4 win their first round games are so

much lower, it is not suprising that $((1,2),(3,4))$ does not maximize the probability the best player wins their first round game. The reason $((1,2),(3,4))$ maximizes the probability that the best player wins the tournament is that players 1 and 2 likely play one easy opponent in the first round and one hard opponent in the second round, as opposed to hard opponents in both rounds. Thus, players 1 and 2 have a higher probability of winning the tournament in bracket $((1,2),(3,4))$ than in bracket $((1,4),(2,3))$ and this increase is enough that the probability the best player wins for bracket $((1,2),(3,4))$ is higher than the probability for bracket $((1,4),(2,3))$.

3.4.2 $N = 8$

For the case when $N = 8$, we focus on two prior distributions, the point mass prior and equal variance prior, and in both cases

$$\mu = (0.21, 0.15, 0.09, 0.03, -0.03, -0.09, -0.15, -0.21), \tag{3.28}$$

corresponding to means that are uniformly spread between -0.21 and 0.21. Interestingly, the optimal brackets for the point mass prior and equal variance prior no longer agree when $N = 8$. Additionally, the optimal bracket depends on the size of the variance.

Point mass prior

For the point mass prior, let

$$\boldsymbol{\theta} = (0.21, 0.15, 0.09, 0.03, -0.03, -0.09, -0.15, -0.21). \tag{3.29}$$

Again, we know player 1 is the best player. The optimal bracket is displayed in Figure 3.10 and the probability player 1 wins the tournament is 0.2318. As before, this bracket only depends on the ordering of the strength parameters and will be optimal as long as $\theta_1 > \theta_2 > \cdots > \theta_8$.

Player 1 plays two of the three easiest teams to advance to the final and once there, will play either player 2, 3, 4, or 5. In the final, the two biggest threats to player 1 are players 2 and 3 and the optimal bracket attempts to eliminate those players before even reaching the final. By matching up players 2 and 3 in the first round, it guarantees that one of them will be eliminated in the first round. Whichever one wins will play either player 4 or 5 in the next round. Ideally, player 1 plays one of the weaker opponents in the final, maximizing his chance of winning the tournament.

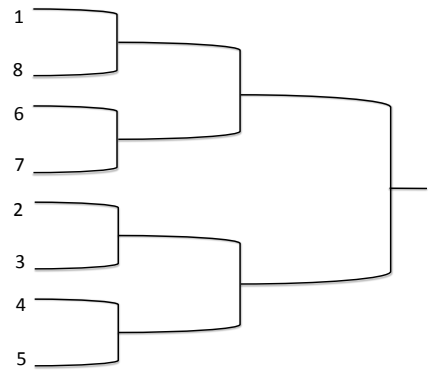


Figure 3.10: **Optimal bracket for point mass prior:** We know player 1 is the best player and the optimal bracket gives player 1 the easiest path to the final. The optimal bracket also gives player 1's biggest rivals, player 2 and 3, the most difficult path to the finals. The probability player 1 wins is 0.2318.

This optimal bracket would likely strike many tournament organizers and players as overly biased in favor of player 1. However, if we know player 1 is the best player and we design the tournament to identify the best player, it is not surprising that we obtain an extreme design. In settings where we do not definitively know the identity of the best player, the optimal brackets are less biased in favor of one player.

Equal variances

We consider two prior distributions, both with equal variance. The optimal brackets differ from each other and both differ from the optimal bracket for the point mass prior case. For the first example, let $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\begin{aligned} \boldsymbol{\mu} &= (0.21, 0.15, 0.09, 0.03, -0.03, -0.09, -0.15, -0.21) \\ \boldsymbol{\Sigma} &= 0.05 \cdot I. \end{aligned} \tag{3.30}$$

The standard deviation is 0.22 and is significantly larger than the differences between consecutive prior means. The probability each player is the best is reported in Table 3.9.

Table 3.9: **Probability each player is the best:** Player 1 has the highest probability of being the best but there is significant probability one of the other players is the best.

Player	Probability player is best
1	0.345
2	0.241
3	0.163
4	0.106
5	0.067
6	0.041
7	0.024
8	0.013

We calcualte $U(\boldsymbol{b})$ for all $m = 315$ brackets and the optimal bracket is shown in Figure 3.11.

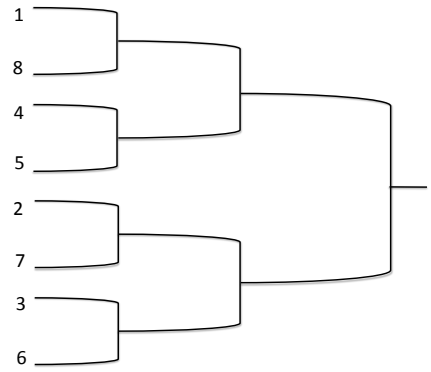


Figure 3.11: **Optimal bracket when $N = 8$ and $\boldsymbol{\Sigma} = 0.05 \cdot I$:** The optimal bracket actually follows the standard seeding. The probability the best player wins is 0.2833. As the mean stays the same but variance increases, for instance when $\boldsymbol{\Sigma} = 0.5 \cdot I$, the optimal bracket continues to follow the standard seeding.

For this bracket, the probability that the best player wins is 0.2833. Note that because player 1 is no longer guaranteed to be the best player, the optimal bracket does not maximize the probability player 1 wins the tournament. For instance, since player 2 is the best player with probability 0.241, this optimal bracket gives player 2 an easier path to win the tournament than the optimal bracket for the point mass prior.

Not all equal variance priors where

$$\boldsymbol{\mu} = (0.21, 0.15, 0.09, 0.03, -0.03, -0.09, -0.15, -0.21) \tag{3.31}$$

have the same optimal bracket. If we change $\boldsymbol{\Sigma}$ from 0.05 to 0.005, the optimal bracket changes to the one in Figure 3.12.

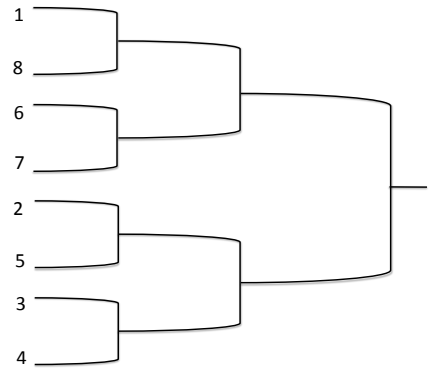


Figure 3.12: **Optimal bracket when $N = 8$ and $\boldsymbol{\Sigma} = 0.005 \cdot I$:** The optimal bracket is nearly equal to the optimal bracket for the point mass prior, reflecting that player 1 is the best with probability 0.67. The probability the best player wins is 0.2284.

For this bracket, the probability the best player wins is 0.2284. Note that this bracket is nearly identical to the optimal bracket for the point mass prior. The difference is that in the optimal bracket for the point mass prior, players 2 and 3 are matched up in the first round. When $\boldsymbol{\Sigma} = 0.005 \cdot I$, player 2 still has probability 0.24 of being the best and thus, player 2 is given an easier matchup in the first round.

3.4.3 $N = 16$

When $N = 16$, the number of brackets is $k = 6.4 \times 10^8$ and the number of win vectors is $m = 32768$, and we turn to the simulated annealing and quasi-Monte Carlo methodology we introduced earlier. Recall that in this case, we only generate an approximately optimal bracket both because simulated annealing is not guaranteed to find the optimum and because quasi-Monte Carlo returns an estimate of $U(\boldsymbol{b})$. We

assess the performance of the approximately optimal bracket by comparing it to the brackets presented Glickman (2008) in terms of the probability that the best player wins. Overall, the approximately optimal brackets perform very well.

In Glickman (2008), the adaptive bracket that repeatedly maximizes the probability the best player advances to the next round is compared to three fixed brackets and one other adaptive bracket. The three fixed brackets are the randomly generated bracket, the bracket following the standard seeding, and the cohort randomized seeding bracket from Schwenk (2000). The adaptive bracket is the reseeding bracket from Hwang (1982). Of those four, all except the random bracket require the players be ordered and we order them according to their prior means. We briefly review the six brackets, five from Glickman (2008) and our optimal fixed bracket.

1. **Random bracket:** The random bracket is a fixed bracket where the players are placed randomly. Each of the k brackets is equally likely.

2. **Standard seeding:** A bracket following the standard seeding for $N = 16$ is shown in Figure 3.13 where the order of the player strengths is assumed to follow the order of the player indices. Thus, player 1 is stronger than player 2, player 2 is stronger than player 3 and so on.

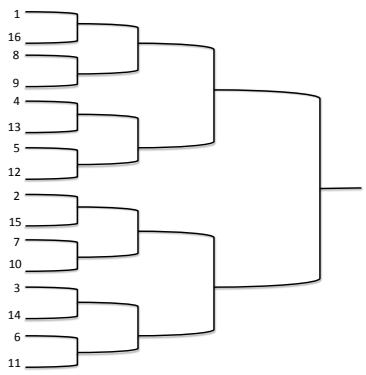


Figure 3.13: **Bracket following standard seeding:** We assume that the player strengths follow the order of the player indices.

3. **Cohort randomized seeding:** Schwenk (2000) introduced cohort randomized seeding, which uses the standard seeding but permutes the player positions within particular groups. We again assume that the order of the player strengths follows the order of the player indices and consider three groups, $\{3, 4\}$, $\{5, 6, 7, 8\}$, and $\{9, 10, 11, 12, 13, 14, 15, 16\}$. The bracket positions of players within each group are randomly permuted. For instance, in group $\{5, 6, 7, 8\}$, the bracket positions of players 5 and 6 may be switched. The resulting bracket is shown in Figure 3.14.

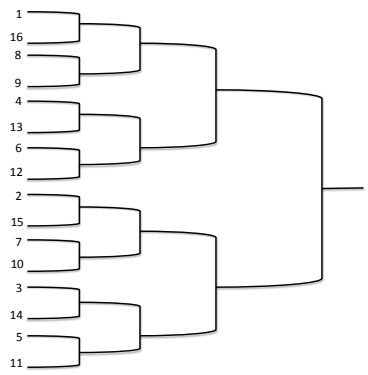


Figure 3.14: **Bracket following cohort randomized seeding:** Because players 5 and 6 are in the same cohort, their positions within the bracket following the standard seeding are permuted.

4. **Reseeding:** Hwang (1982) proposed reseeding the players after each round. For instance, say there are $n = N/2^r$ players remaining after round r. The players are ordered according to their prior means and the ith best player is matched up with the $(n - i + 1)$th best player. In the standard bracket in Figure 3.13, if player 16 defeats player 1 and player 2 defeats player 15 in the first round, player 2 would play player 16 in the second round, rather than the winner of the game between players 7 and 10.

5. **Best player (BP) advances:** As we have mentioned, Glickman (2008) assumes a multivariate normal prior on $\boldsymbol{\theta}$ and for each round, finds the matchups that maximize the probability the best player advances to the next round. This is an adaptive bracket and the utility function is local in the sense that we are maximizing the probability the best player advances to the next round, not the

probability the best player wins the tournament.

6. **BP wins:** Finally, we include the fixed bracket that maximizes the probability the best player wins. We use the simulated annealing parameters discussed previously with 3000 quasi-Monte Carlo draws. The running time for one simulated annealing chain is approximately 2 minutes on a 3.2GHz quad-core CPU.

We consider the six different prior distributions (A-F) from Glickman (2008).

(A) $\mu_i = 0.75 - 0.1(i-1)$; $\sigma_i^2 = 0.1$ for all i.

(B) $\mu_i = 0.75 - 0.1(i-1)$; $\sigma_i^2 = 0.01$ for odd i, and $\sigma_i^2 = 1.0$ for even i.

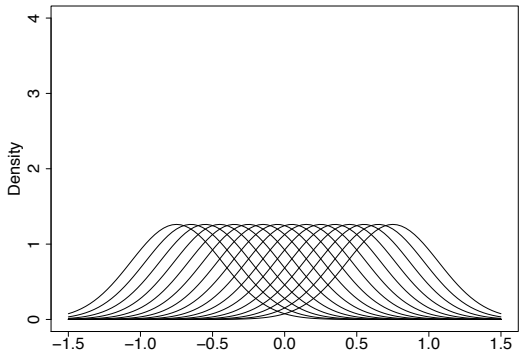
(C) $\mu_i = 0.75 - 0.1(i-1)$; $\sigma_i^2 = 0.01$ for $i \leq 8$, and $\sigma_i^2 = 1.0$ for even $i \geq 9$.

(D) $\mu_i = 0.75 - 0.1(i-1)$; $\sigma_i^2 = 1.0$ for $i \leq 8$, and $\sigma_i^2 = 0.01$ for even $i \geq 9$.

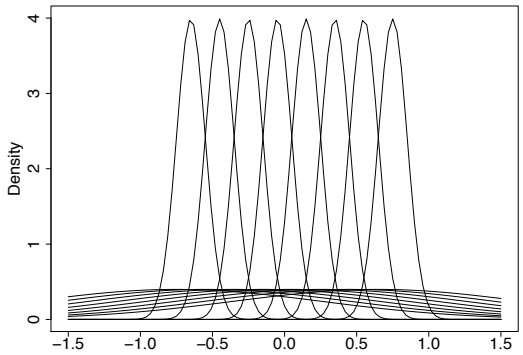
(E) $\mu_i = 0.75 - 0.1(i-1)$; $\sigma_i^2 = 0.01$ for $i \leq 4$, and $\sigma_i^2 = 1.0$ for even $i \geq 5$.

(F) $\mu_i = 0.75 - 0.1(i-1)$ for $i = 1, \ldots, 12$, $\mu_i = -0.60 - 0.05(i-13)$ for $i = 13, \ldots, 16$; $\sigma_i^2 = 0.01$ for $i \leq 8$, and $\sigma_i^2 = 1.0$ for even $i \geq 9$.

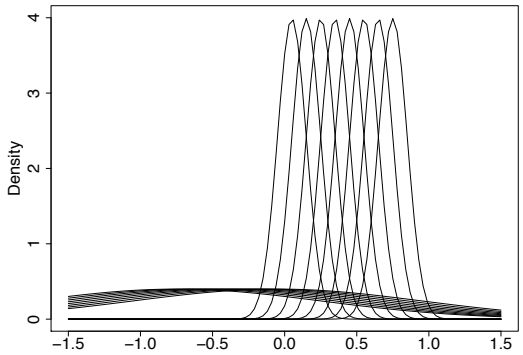
We visualize the marginal distribution of θ_i for the six prior distributions in Figure 3.15.



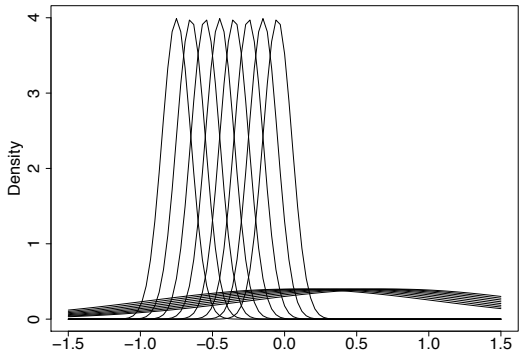
(a) Prior A



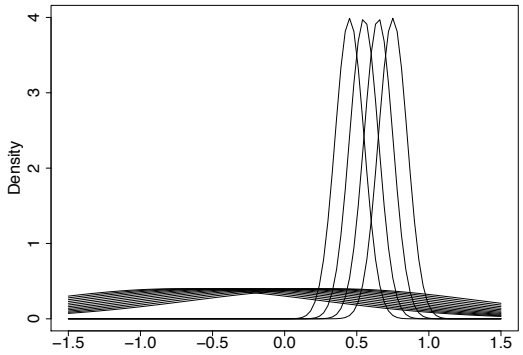
(b) Prior B



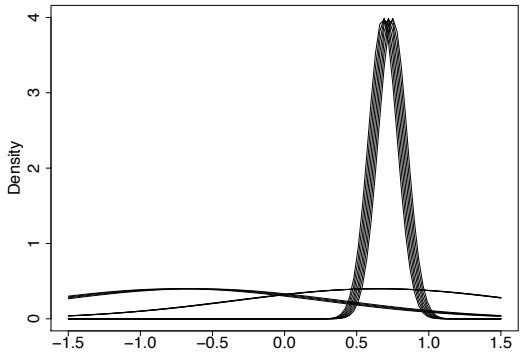
(c) Prior C



(d) Prior D



(e) Prior E



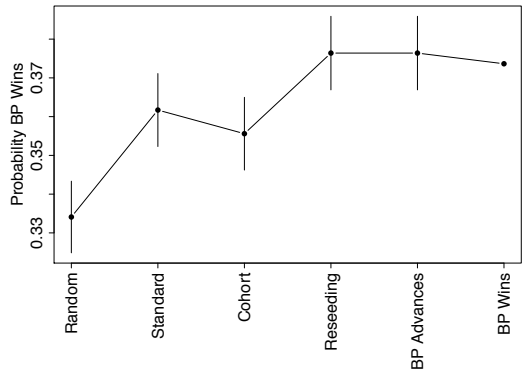
(f) Prior F

Figure 3.15: **Six marginal prior distributions**

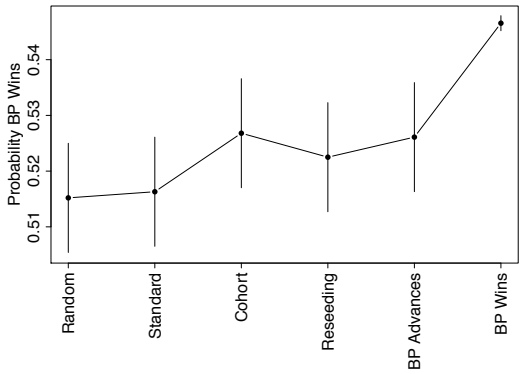
In Table 3.10, we report the probability the best player wins for all combinations of the six brackets and six prior distributions. We have copied the results from Glickman (2008) for the five brackets he considered. See Glickman (2008) for further details regarding the simulation set-up. For the BP wins bracket and for each prior distributions, we draw a value of $\boldsymbol{\theta}$ and then simulate the 15 games of the tournament. We record whether the best player, the player with the largest θ_i, won and repeat this process 100000 times. The last column in Table 3.10 reports the average number of times the best player won. The results are also reported graphically in Figure 3.16, which also includes confidence intervals.

Table 3.10: **Probability best player wins tournament for different brackets and priors:** The results in the first five columns are copied from Glickman (2008). The last column was found by simulation 100000 tournaments for the approximately optimal brackets.

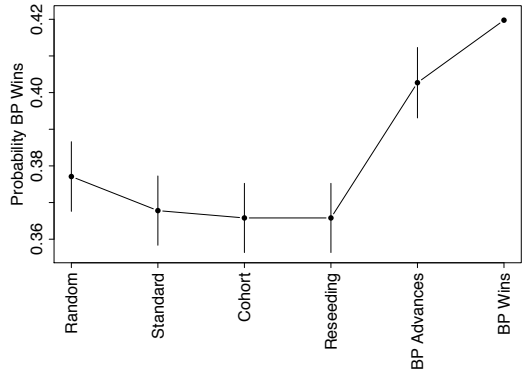
Prior	Random	Standard	Cohort	Reseeding	BP Advances	BP Wins
A	0.3341	0.3617	0.3556	0.3764	0.3764	0.3728
B	0.5152	0.5163	0.5268	0.5225	0.5261	0.5446
C	0.2771	0.3678	0.3658	0.3685	0.4027	0.4207
D	0.5852	0.5924	0.6003	0.5979	0.5972	0.5974
E	0.4710	0.4620	0.4571	0.4633	0.4781	0.4994
F	0.4692	0.4548	0.4525	0.4542	0.5064	0.5179



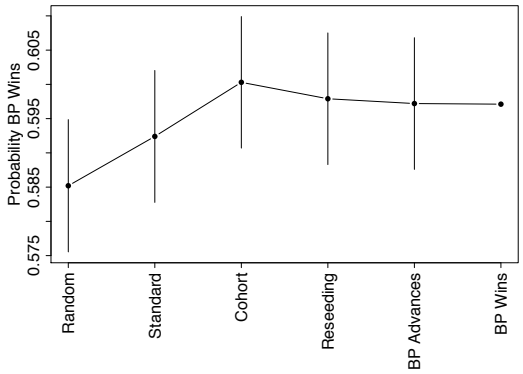
(a) Prior A



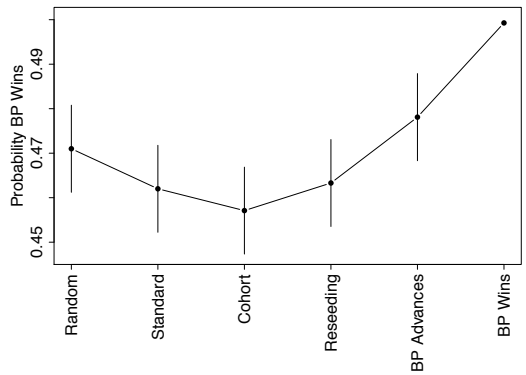
(b) Prior B



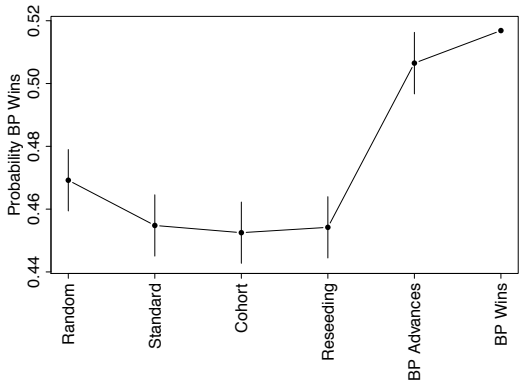
(c) Prior C



(d) Prior D



(e) Prior E



(f) Prior F

Figure 3.16: **Probability best player wins tournament**

Overall, the BP wins bracket is very competitive in relation to the other five brackets, including the two adaptive brackets. In 3 of the 6 prior distributions (priors B, C, and E), the probability the best player wins is clearly the highest for the BP wins bracket. For prior F, the BP wins bracket also likely gives the largest probability. These four priors are also the priors for which the BP advances bracket performs well. Glickman (2008) noted that for these priors, the "top players' strengths are precisely estimated, and the bottom players are imprecisely estimated." He also notes that

> In gaming organizations, it is often the case that the best players compete more frequently than weaker players and therefore have strengths that are more precisely estimated, so that our pairing method would be ideal for such a scenario.

This conclusion applies to the BP wins bracket as well. For priors A and D, many of the brackets give similar results and the BP wins bracket is likely equal to the other brackets.

It is somewhat surprising that the BP wins bracket gives consistently higher probabilities than the BP advances bracket because the BP advances bracket is adaptive and has greater flexibility. However, as we have seen, maximizing the probability the best player advances is not equivalent to maximizing the probability the best player wins the tournament. The results highlight that we can achieve a higher probability of the best player winning the tournament by using the correct utility function and a more constrained bracket than using an incorrect utility function and a more flexible bracket.

3.5 Other utility functions

There are many utility functions we could consider besides the best player winning the tournament. In what follows, we consider three other utility functions. Each one extends the focus beyond the best player. The first utility function we consider looks at the best two players and the next two utility functions are defined in terms of all N players. Other utility functions that might also be of interest include those related to the "entertainment" value of the games. For instance, we might be interested in maximizing the number of games between evenly matched players because those games are more exciting for the audience.

3.5.1 Two best players meet in the final

Tournament organizers often want the best players to meet in the later rounds, when the stakes are higher. We can adjust the utility function, for instance, to maximize the probability the two best players meet in the final. Let the utility function be

$$u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}) = I(w_{(1)}(\boldsymbol{\theta}) \geq R-1,\, w_{(2)}(\boldsymbol{\theta}) \geq R-1). \tag{3.32}$$

The expected utility is then

$$U(\boldsymbol{b}) = \int_{\Theta} \mathrm{p}(w_{(1)}(\boldsymbol{\theta}) \geq R-1,\, w_{(2)}(\boldsymbol{\theta}) \geq R-1 \,|\, \boldsymbol{\theta}, b)\, \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{3.33}$$

We can slightly adapt the methodology developed previously to find the optimal bracket in this case. We return to our example when $N = 8$ and $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (0.21, 0.15, 0.09, 0.03, -0.03, -0.09, -0.15, -0.21)$$
$$\boldsymbol{\Sigma} = 0.005 \cdot I. \tag{3.34}$$

The optimal bracket is shown in Figure 3.17.

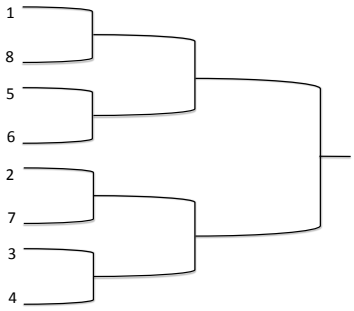


Figure 3.17: **Optimal bracket when $N = 8$ for two best players meeting in final when $\Sigma = 0.005 \cdot I$:** The probability the two best meet in the final is 0.1224. Note that player 2 is now matched up with player 7.

Note that when maximizing the probability the best player wins, player 2 was matched up with player 5 in the first round, see Figure 3.12. Now, player 2 is matched up in the first round against player 7, an easier match up. This is not surprising since players 1 and 2 are most likely the two best players.

We also find the bracket that approximately maximizes the probability the two best players meet in the final when $N = 16$. We let $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\begin{aligned}\boldsymbol{\mu} &= (0.45, 0.39, 0.33, 0.27, 0.21, 0.15, 0.09, 0.03,\\ &\quad -0.03, -0.09, -0.15, -0.21, -0.27, -0.33, -0.39, -0.45)\\ \boldsymbol{\Sigma} &= 0.005 \cdot I. \end{aligned} \tag{3.35}$$

The optimal bracket is shown in Figure 3.18.

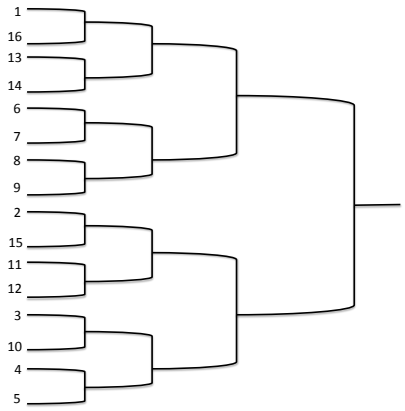


Figure 3.18: **Approximately optimal bracket when $N = 16$ for two best players meeting in final when $\Sigma = 0.005 \cdot I$:** The probability the two best players meet in the final is 0.1112.

3.5.2 w is a monotonic function of θ

The tournament organizer could also be interested in the best four players meeting in the semi-finals or the best 8 players meeting in the quarter-finals. These utility functions can actually be combined. For instance, we can maximize the probability that the best player wins, the best two players meet in the final, the best four players

meet in the semi-finals, and so on. Taken to the extreme, we could maximize the probability that $\boldsymbol{w}$ is a monotonic function of $\boldsymbol{\theta}$. We can also think of this as maximizing the probability each player wins the "correct" number of games, where by correct number of games, we mean that the worst $N/2$ players win 0 games, the next worst $N/4$ players win 1 game, the next worst $N/8$ players win two games and so on. The corresponding utility function is

$$u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}) = I(w_{(i)}(\boldsymbol{\theta}) = R - \lceil \log_2(i) \rceil \ \forall i), \tag{3.36}$$

and the expected utility function is

$$U(\boldsymbol{b}) = \int_{\Theta} \mathrm{p}(w_{(i)}(\boldsymbol{\theta}) = R - \lceil \log_2(i) \rceil \ \forall i \,|\, \boldsymbol{\theta}, b)\, \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{3.37}$$

For the case when $N = 8$ and $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\begin{aligned} \boldsymbol{\mu} &= (0.21, 0.15, 0.09, 0.03, -0.03, -0.09, -0.15, -0.21) \\ \boldsymbol{\Sigma} &= 0.005 \cdot I, \end{aligned} \tag{3.38}$$

the bracket that maximizes the probability that $\boldsymbol{w}$ is a monotonic function of $\boldsymbol{\theta}$ is shown in Figure 3.19.

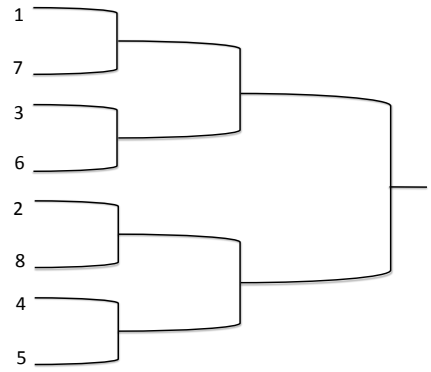


Figure 3.19: **Optimal bracket when $N = 8$ for probability $\boldsymbol{w}$ is a monotonic function of $\boldsymbol{\theta}$ when $\Sigma = 0.005 \cdot \boldsymbol{I}$:** The probability each player wins the correct number of games is 0.0111.

Interestingly, this bracket gives player 2 an easier path to the final than player 1. Because we are maximizing the probability that $\boldsymbol{w}$ is a monotonic function of $\boldsymbol{\theta}$, it appears that the easier path to the finals helps player 2 more than a slightly harder path hurts player 1. Also, by having player 1 likely play player 3 in the second round, we increase the probability player 3 correctly wins one game rather than two.

When $N = 16$, finding the optimal bracket for such a utility function requires sampling from $\mathscr{W}_{\boldsymbol{b}}$ because sampling from Θ is inefficient. For a given bracket, say we sample $\boldsymbol{\theta}$ from Θ. For that $\boldsymbol{\theta}$, in order for the probability that $\boldsymbol{w}$ is a monotonic function of $\boldsymbol{\theta}$ not to be 0, none of the $N/2$ best players can play each other in the first round, none of the $N/4$ best players can play each other in the second round, and so on. In general, this is unlikely for a random $\boldsymbol{\theta}$ draw. This is not an issue when sampling from $\mathscr{W}_{\boldsymbol{b}}$ but sampling from $\mathscr{W}_{\boldsymbol{b}}$ is much slower. Finding the optimal

bracket would take several days on a standard personal computer and consequently, we leave the implementation of this method for future work. It is also possible that related utility functions, such as the rank correlation between $\boldsymbol{w}$ and $\boldsymbol{\theta}$, could be more manageable computationally but result in the same or nearly the same optimal bracket. We also leave this for future work.

3.5.3 Standard seeding

The idea that $\boldsymbol{w}$ is a monotonic function of $\boldsymbol{\theta}$ is closely related to the standard seeding. In the standard seeding, if the better teams always wins, the worst $N/2$ players win 0 games, the next worst $N/4$ players win 1 game, the next worst $N/8$ players win two games and so on. The standard seeding also has the appealing property that, in general, the better the team, the easier the path to win the tournament. Another possible utility function is maximizing the probability that the bracket follows the standard seeding. This is potentially useful to tournament organizers who are understandably attached to the familiar standard seeding and want to account for uncertainty in player strength.

When the prior variances for each player are the same, there is a natural ordering of the competitors and applying the standard seeding is straightforward. However, when the prior variances are different, there is not a natural ordering of the players. As we have seen, the player with the largest prior mean can also have the largest variance, in which case, that player can both have the largest probability of being the best player and the largest probability of being the worst player. Tournament organizers may be hesitant to assign such a player the top seed in the tournament.

We return to one of our examples where $N = 4$. Let

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} \sim N\left(\begin{pmatrix} 0.09 \\ 0.03 \\ -0.03 \\ -0.09 \end{pmatrix}, \begin{pmatrix} 1.0 & 0 & 0 & 0 \\ 0 & 1.0 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0.01 \end{pmatrix} \right). \tag{3.39}$$

In each of the following eight orderings of the 4 players, the bracket $((1,2),(3,4))$ follows the standard seeding.

$$\begin{aligned} &\theta_1 > \theta_3 > \theta_4 > \theta_2 \\ &\theta_1 > \theta_4 > \theta_3 > \theta_2 \\ &\theta_2 > \theta_3 > \theta_4 > \theta_1 \\ &\theta_2 > \theta_4 > \theta_3 > \theta_1 \\ &\theta_3 > \theta_1 > \theta_2 > \theta_4 \\ &\theta_3 > \theta_2 > \theta_1 > \theta_4 \\ &\theta_4 > \theta_1 > \theta_2 > \theta_3 \\ &\theta_4 > \theta_2 > \theta_1 > \theta_3 \end{aligned} \tag{3.40}$$

We let $\mathscr{S}_{\boldsymbol{b}} \subset \Theta$ be the set of $\boldsymbol{\theta}$ values that satisfy the standard seeding for bracket $\boldsymbol{b}$. Each $\boldsymbol{\theta} \in \mathscr{S}_{\boldsymbol{b}}$ satisfies one of the orderings. The sets $\{\mathscr{S}_{\boldsymbol{b}} : \boldsymbol{b} \in \mathscr{B}\}$ partition Θ since every $\boldsymbol{\theta}$ follows the standard seeding for exactly one bracket. We then let the utility function be

$$u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}) = I(\boldsymbol{\theta} \in \mathscr{S}_{\boldsymbol{b}}). \tag{3.41}$$

and the expected utility is

$$\begin{aligned} U(\boldsymbol{b}) &= \int_{\Theta} \sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} u(\boldsymbol{b}, \boldsymbol{w}, \boldsymbol{\theta}) \, \mathrm{p}(\boldsymbol{w} \mid \boldsymbol{\theta}, \boldsymbol{b}) \, \mathrm{p}(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \qquad (3.42) \\ &= \int_{\Theta} \sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} I(\boldsymbol{\theta} \in \mathscr{S}_{\boldsymbol{b}}) \, \mathrm{p}(\boldsymbol{w} \mid \boldsymbol{\theta}, \boldsymbol{b}) \, \mathrm{p}(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \\ &= \int_{\Theta} I(\boldsymbol{\theta} \in \mathscr{S}_{\boldsymbol{b}}) \sum_{\boldsymbol{w} \in \mathscr{W}_{\boldsymbol{b}}} \mathrm{p}(\boldsymbol{w} \mid \boldsymbol{\theta}, \boldsymbol{b}) \, \mathrm{p}(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \\ &= \int_{\Theta} I(\boldsymbol{\theta} \in \mathscr{S}_{\boldsymbol{b}}) \, \mathrm{p}(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \\ &= \int_{\mathscr{S}_{\boldsymbol{b}}} \mathrm{p}(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \\ &= \Pr(\boldsymbol{\theta} \in \mathscr{S}_{\boldsymbol{b}}) \qquad (3.43) \end{aligned}$$

Note that this utility function does not involve $\boldsymbol{w}$ because in this case, the utility function is not related to the outcome of the tournament. The expected utility is then an integral of $\mathrm{p}(\boldsymbol{\theta})$ over $\mathscr{S}_{\boldsymbol{b}}$. However, it is easier to carry out the integral by separately integrating over the $\boldsymbol{\theta}$ values that satisfy each order. For instance, for the 4-player bracket $((1,2),(3,4))$, we sum the integrals of $\mathrm{p}(\boldsymbol{\theta})$ over the 8 orders, such that

$$\begin{aligned}\int_{\mathscr{S}_{\boldsymbol{b}}} \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = &\int_{\theta_1>\theta_3>\theta_4>\theta_2} \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} + \int_{\theta_1>\theta_4>\theta_3>\theta_2} \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &+ \int_{\theta_2>\theta_3>\theta_4>\theta_1} \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} + \int_{\theta_2>\theta_4>\theta_3>\theta_1} \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &+ \int_{\theta_3>\theta_1>\theta_2>\theta_4} \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} + \int_{\theta_3>\theta_2>\theta_1>\theta_4} \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &+ \int_{\theta_4>\theta_1>\theta_2>\theta_3} \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta} + \int_{\theta_4>\theta_2>\theta_1>\theta_3} \mathrm{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \end{aligned} \tag{3.44}$$

For each bracket, $\boldsymbol{b}$, there are $\frac{N!}{k} = 2^{N-1}$ orderings that satisfy the standard seeding and the sum of these probabilities equals $\Pr(\boldsymbol{\theta} \in \mathscr{S}_b)$. To calculate $U(\boldsymbol{b})$ for each bracket in the 4-player case, we have to carry out $4! = 24$ integrals, 8 integrals per bracket. The probability of the standard seeding for each of the three brackets is reported in Table 3.11.

Table 3.11: **Probability of standard seeding for 3 brackets:** Bracket $((1,2),(3,4))$ has the highest probability of satisfying the standard seeding.

Bracket	Prob. of standard seeding
$((1,2),(3,4))$	0.4512
$((1,3),(2,4))$	0.2723
$((1,4),(2,3))$	0.2763

Bracket $((1,2),(3,4))$ maximizes the probability of the standard seeding. Note that players 1 and 2 have both the highest probability of being the best and the highest probability of being the worst. Thus, matching them up in the first round maximizes the probability that the best player plays the worst player in the first round, which

is required by the standard seeding.

We also consider an example where $N = 8$ and $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\begin{aligned} \boldsymbol{\mu} &= (0.21, 0.15, 0.09, 0.03, -0.03, -0.09, -0.15, -0.21) \\ \boldsymbol{\Sigma} &= \text{diag}(1.0, 1.0, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01). \end{aligned} \tag{3.45}$$

When $N = 8$, we have to carry out $8! = 40320$ integrals, 128 integrals per bracket. These 40320 integrals took approximately 6 minutes on a 3.2GHz quad-core CPU. The optimal bracket is shown in Figure 3.20.

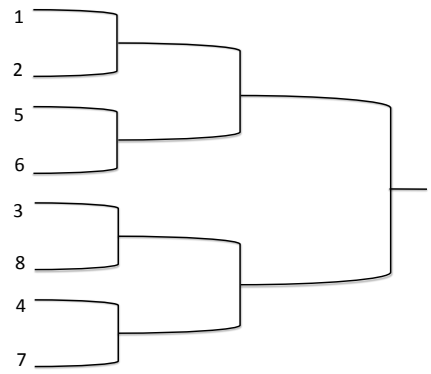


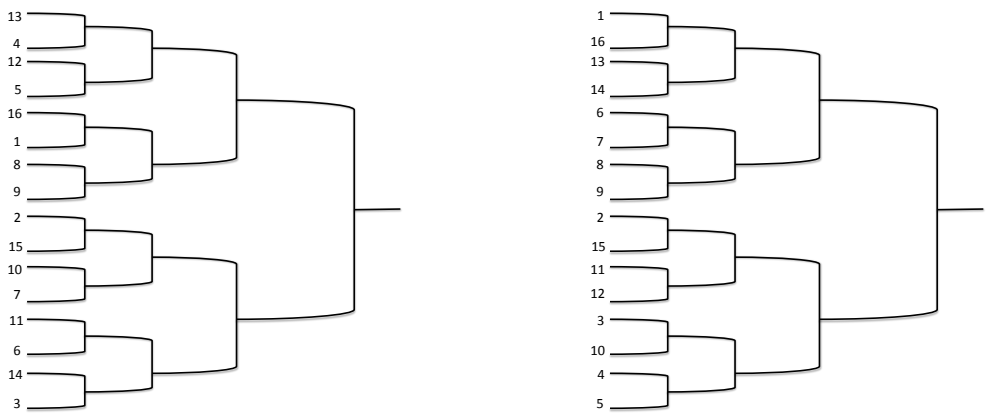
Figure 3.20: **Optimal 8-player bracket for standard seeding:** The probability of the standard seeding is 0.0088.

We see a similar result when $N = 8$. The optimal bracket also matches up the two players with the largest variance because this matchup has the best chance of being between the best team and the worst team.

When $N = 16$, there are $16! \approx 2 \cdot 10^{13}$ total orderings and 32768 orderings per bracket. We again turn to our simulated annealing approach. Calculating $U(\boldsymbol{b})$ for a single bracket requires carrying out 32768 16-dimensional integrals and takes approximately 30 minutes. Consequently, we must estimate $U(\boldsymbol{b})$ and one natural estimation approach is, for each bracket $\boldsymbol{b}$, to draw a random sample of L orderings from the total number of 2^{N-1} orderings that satisfy the standard seeding. Generating these orderings can be done by generating brackets like the one in Figure 3.21(a), where the player labels refer to the rank of the players, not the player indices. Although written in an unusual form, this is an example of a bracket that follows the standard seeding. We can combine it with a bracket, such as the one in Figure 3.21(b), where the labels refer to the player indices. For instance, the two brackets in Figure 3.21, imply that an ordering of

$$\theta_7 > \theta_2 > \theta_5 > \theta_{16} > \theta_{14} > \theta_{10} > \theta_{12} > \theta_8 > \theta_9 > \theta_{11} > \theta_3 > \theta_{13} > \theta_1 > \theta_4 > \theta_{15} > \theta_6 \tag{3.46}$$

would satisfy the standard seeding for the bracket in 3.21(b).



(a) Labels correspond to player ranks (b) Labels correspond to player indices

Figure 3.21: **How to find orderings that satisfy the standard seeding:** The two brackets can be combined to generate an ordering that follows the standard seeding for the bracket in (b).

Brackets like the one in Figure 3.21(a) can be randomly generated by swapping players, games, and sets of games in such a way as to preserve the standard seeding. We generate a series of such brackets and then thin the series to reduce autocorrelation. We then take the mean of the l probabilities associated with each ordering and multiply the result by 2^{N-1} to estimate $\Pr(\boldsymbol{\theta} \in \mathscr{S}_{\boldsymbol{b}})$. However, even carrying out 1000 integrals takes approximately 1 minute and a sample of 1000 draws does not provide enough precision. We leave speeding up this calculation for future work.

3.6 Conclusion

We have presented a methodology for finding optimal fixed knockout tournament brackets when player strengths follow a prior distribution. This work extends Glick-

man (2008) by considering utility functions that apply to players' tournament results rather than just one game. We focused on maximizing the probability that the best player wins the tournament, a historically important objective that is especially relevant when we are uncertain which player is the best. We found that the approximately optimal bracket outperforms many of its competitors, including the adaptive brackets. Finally, we considered alternative utility functions that may be appealing to some tournament organizers.

Chapter 4

Inference for causal effects in 2^2 factorial experiments with non-compliance

4.1 Introduction

Factorial experiments are among the most popular experimental designs because they allow for the effects of multiple treatment factors and the interactions between them to be estimated simultaneously and in an efficient manner. They were originally developed in the context of agricultural experiments (Yates, 1937; Fisher, 1935) and, at the time, the ability to estimate multiple effects simultaneously was not generally accepted. To quote Fisher (1926),

> No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question, at a time. The writer is convinced that this view is wholly mistaken.

History has proved Fisher correct and factorial experiments are now used extensively in industrial and medical experiments.

Each factor in a factorial experiments has a certain number of levels. The number of levels is typically two although any number is possible. For example, in a medical trial, one of the factors could be aspirin use and the two levels could be taking aspirin (level 1) or not taking aspirin (level 0). In a factorial experiment, every possible treatment combination, a combination of the factor levels, is tested at least once. If there are k factors and each factor has two levels, then all 2^k treatment combinations are tested and the experiment is said to follow a 2^k factorial design. Say there are $k = 2$ factors and that the two factors are aspirin use and beta-carotene use. Then the $2^2 = 4$ treatment combinations are (no aspirin, no beta-carotene), (no aspirin, beta-carotene), (aspirin, no beta-carotene), and (aspirin, beta-carotene) (Stampfer et al., 1985).

Factorial designs have been widely applied in agricultural, industrial, and medical contexts and this project was motivated by an experiment in education. The New York City Department of Education was interested in simultaneously evaluating the effect of multiple school initiatives on student performance. The initiatives included linking teacher bonuses to student performance and a new web-based student tracking system called ARIS. Schools would be randomly assigned to one of the treatment combinations and student performance would be measured at the end of the year. While the experiment never moved beyond the design stage, it raised important questions. For instance, one of the complications of the design was that the Department of Education did not want to force a subset of schools to adopt the ARIS system. As a

result, a school could be randomly assigned to use the ARIS system but choose not to use it. If we are interested in the effect of actually using the ARIS system, this is a problem because the non-compliance breaks the randomization. Randomization ensures that, on average, the schools assigned to use the ARIS system are identical to the schools assigned not to use the ARIS system and thus, these two sets of schools allow for a fair comparison. However, if schools are not forced to comply with their assignment, there is no guarantee that the schools that use the system are similar to the schools that do not. This can also be viewed as an example of an encouragement design (Hirano et al., 2000). In this chapter, we will focus on estimating factorial effects in the presence of such non-compliance.

Non-compliance is a well known problem in experimental design and more generally. Steiner (2012) estimates that "fewer than 50% of individuals prescribed a new medication for diabetes, hypertension, or hyperlipidemia continue the drug for even a year." Non-compliance has also been extensively studied in the statistics literature (Angrist et al., 1996; Imbens and Rubin, 1997; Cheng and Small, 2006; Jin and Rubin, 2008; Roy et al., 2008; Little et al., 2009; Long et al., 2010). The principal stratification approach, introduced by Frangakis and Rubin (2002), has been especially effective for such problems. In principal stratification, each unit belongs to a single strata and because this strata is determined before treatment assignment, it is a proper covariate. The goal is to compute causal effects within each strata. However, the strata identifiers are unknown variables and we have to use the observed compliance behavior and outcomes to infer which unit belongs to which strata. In a non-compliance setting, the strata are typically whether or not the units are compli-

ers, never-takers, always-takers, or defiers, terms that will be defined more formally in the next section. Most of the literature has focused on experiments with a single factor with two levels. While Cheng and Small (2006), Roy et al. (2008) and Long et al. (2010) extended the principal stratification framework to single factor experiments with three levels, there has been limited work on how to apply the framework to factorial experiments.

We focus primarily on the 2^2 experiment and, given certain assumptions, the framework enables the estimation of both factorial main effects and the interaction. The work can be viewed as a combination of the potential outcomes framework for factorial experiments presented in Dasgupta et al. (2012) and principal stratification.

In Section 4.2, we state the problem of non-compliance in factorial experiments more formally and lay out the key assumptions. In Section 4.3, we present the principal stratification estimation strategy and in Section 4.4, present simulation results. In Section 4.5, we present two extensions and in Section 4.6 we summarize our findings.

4.2 The problem

Let F_1 and F_2 be two two-level treatment factors in a 2^2 factorial design. Additionally, let N be the total number of units, where $N/4$ units are randomly assigned to each treatment combination. As an illustrative example, we consider the factors from Bhasin et al. (1996). In this study, F_1 was a standardized weight lifting program to be performed three times a week and F_2 was a weekly steroid injection. The goal of the 10-week study was to assess the effects of the weight lifting program, the steroids, and the interaction between them on strength, as measured by muscle size and bench-press

weight. In what follows, we assume that the subjects either comply fully with their assignment or not at all (i.e. all-or-nothing compliance). For instance, the subjects either completely follow the standardized weight lifting program or completely ignore it. Whether or not this is a realistic simplifying assumption depends on the context and if it is more appropriate, we could re-define compliance to be attending at least 80% of the weight lifting sessions. Also, see Jin and Rubin (2008) for an example of how partial compliance can he handled in the principal stratification framework. We next define the principal strata, assumptions, estimands, and observed data.

4.2.1 Principal strata

Let Z_1 be the assigned level of factor F_1 and let Z_2 be the assigned level of factor F_2. Let $Z_1 = 1$ denote the active level of factor F_1 and let $Z_1 = 0$ denote the control level. For instance, if F_1 refers to the weight lifting program, $Z_1 = 1$ refers to being assigned to the program and $Z_1 = 0$ refers to not being assigned to the program. Then let $Z = (Z_1, Z_2)$ denote the assigned treatment combination. Let $\mathscr{F}$ be the set of treatment combinations,

$$\mathscr{F} = \{(0,0), (0,1), (1,0), (1,1)\}. \tag{4.1}$$

Let $W_i(Z) = W_i(Z_1, Z_2) = (W_{i1}(Z_1, Z_2), W_{i2}(Z_1, Z_2))$ be the received treatment combination for the ith unit when assigned to treatment combination Z. $W_{i1}(Z_1, Z_2)$ is the received level of F_1 and $W_{i2}(Z_1, Z_2)$ is the received level of F_2. Note that $W_i(\cdot)$ is a function from $\mathscr{F}$ to $\mathscr{F}$, $W_i(\cdot) : \mathscr{F} \to \mathscr{F}$. Let

$$\begin{aligned}\boldsymbol{W}_i &= (W_i(Z))_{Z\in\mathscr{F}} \\ &= (W_i(0,0), W_i(0,1), W_i(1,0), W_i(1,1))\end{aligned} \tag{4.2}$$

be the vector of received treatment combinations for the ith unit. The vector of received treatment combinations, $\boldsymbol{W}_i$, defines the function $W_i(\cdot)$ and the compliance behavior for the ith unit. We implicitly make the stable unit treatment value assumption (SUTVA (Rubin, 1980)), which implies that the received treatment combination for unit i only depends on the assigned treatment combination for unit i. Let $\boldsymbol{W}$ be the $N \times 4$ matrix of received treatment combinations where the ith row is $\boldsymbol{W}_i$.

In this context, a principal strata is a subset of the N units with the same value of $\boldsymbol{W}_i$. In the 2^2 design, $|\mathscr{F}| = 4$ and thus $\boldsymbol{W}_i$ take on one of $4^4 = 256$ possible values. There are then 256 principal strata in a 2^2 design and $(2^k)^{(2^k)}$ principal strata in a 2^k design.

4.2.2 Assumptions

In order to make estimation feasible, we reduce the number of principal strata through a series of four assumptions.

1. **No compliance interaction:** We assume there are no interactions among factors with respect to compliance behavior. This implies that $W_{i1}(Z_1, 1) = W_{i1}(Z_1, 0)$ and $W_{i2}(1, Z_2) = W_{i1}(0, Z_2)$. Thus, the received level of factor F_1 only depends on Z_1, and we write $W_{i1}(Z_1, Z_2)$ as $W_{i1}(Z_1)$. Similarly, we write

the received level for F_2 as $W_{i2}(Z_2)$. Under this assumption, with respect to F_j, a unit can be either a complier ($W_{ij}(Z_j) = Z_j$), never-taker ($W_{ij}(Z_j) = 0$), always-taker ($W_{ij}(Z_j) = 1$), or defier ($W_{ij}(Z_j) = 1 - Z_j$). Thus, for a 2^2 design, there are then $4^2 = 16$ principal strata and for a 2^k design, there are 4^k principal strata.

2. **Monotonicity:** We assume for each factor that the received level is no less than the assigned level.

$$\begin{aligned} F_1 : W_i(1, Z_2) &\geq W_i(0, Z_2) \; \forall \, Z_2 \\ F_2 : W_i(Z_1, 1) &\geq W_i(Z_1, 0) \; \forall \, Z_1 \end{aligned} \tag{4.3}$$

This rules out the defiers, who always receive the opposite of their assinged level, and reduces the number of principal strata to $3^2 = 9$ for a 2^2 design and 3^k for a 2^k design.

3. **Strict compliance for F_2:** We assume that units always comply with the assignment of F_2. In the steroid example, this implies that all subjects comply with their assignment either to receive a weekly steroid injection or to not receive the injection. This further reduces the number of principal strata to 3. In general, if there are k factors and strict-compliance holds for m factors, then given the other two assumptions, there are 3^{k-m} principal strata.

With these three assumptions, we have dramatially reduced the number of principal strata from 256 to 3, compliers, never-takers, and always-takers with respect to F_1. Let $C_i = c$ if unit i is a complier, let $C_i = n$ if unit i is a never-taker, and let $C_i = a$ if the unit i is an always-taker. Let $\boldsymbol{C} = (C_1, \ldots, C_N)$ and let $\mathcal{C}(t) = \{i | C_i = t\}$ for $t \in \{c, n, a\}$ be the collection of units in principal strata t. Let $N_t = |\mathcal{C}(t)|$ be number of units in principal strata t and let $p_t = N_t/N$ be the proportion of units in principal strata t.

We next introduce notation for the outcome of interest, Y. In the steroid example, Y would be muscle mass. Similar to how $W_i(Z)$ is the received treatment combination for treatment assignment Z, let $Y_i(Z, W_i(Z))$ be the potential outcome for treatment assignment Z. Since $W_i(Z)$ is a function of Z, we could write $Y_i(Z, W_i(Z))$ as a function of only Z but we include $W_i(Z)$ for notational convenience. In the same way that $\boldsymbol{W}_i = (W_i(0,0), W_i(0,1), W_i(1,0), W_i(1,1))$ is the vector of received treatment combinations for the possible treatment assignments, let

$$\begin{aligned}\boldsymbol{Y}_i &= \big(\big(Y_i(Z, W_i(Z))\big)\big)_{Z \in \mathscr{Z}} \\ &= (Y_i((0,0), W_i(0,0)), Y_i((0,1), W_i(0,1)), Y_i((1,0), W_i(1,0)), Y_i((1,1), W_i(1,1))\end{aligned} \tag{4.4}$$

be the vector of the potential outcomes for the possible treatment assignments. Let $\boldsymbol{Y}$ be the $N \times 4$ matrix of potential outcomes where the ith row is $\boldsymbol{Y}_i$. We make one more assumption before defining the estimands.

4. **Weak exclusion restriction:** We assume that for never-takers and always

takers, the assigned treatment for F_1 is unrelated to the potential outcomes for Y. That is for $i \in \mathcal{C}(n)$ and $i \in \mathcal{C}(a)$

$$\begin{aligned} Y_i((0,0), W_i(0,0)) &= Y_i((1,0), W_i(1,0)) \\ Y_i((0,1), W_i(0,1)) &= Y_i((1,1), W_i(1,1)). \end{aligned} \tag{4.5}$$

The weak exclusion restriction says that for never-takers and always-takers the potential outcomes only depend on the received treatment combination. Since the assigned treatment combination and the received treatment combination are the same for compliers, this implies that under the weak exclusion restriction, we can write $Y_i(Z, W_i(Z))$ as $Y_i(W_i(Z))$.

4.2.3 Estimands

The unit-level Intention-To-Treat (ITT) factorial effects are defined as follows. Let θ_{ij}^{ITT} be the ITT effect for unit i and and factorial effect j. In a 2^k factorial design, there $2^k - 1$ factorial effects. Consequently, in a 2^2 factorial design there are three factorial effects, the two main effects for the two factors and the interaction. Let

$$\theta_{ij}^{\text{ITT}} = \frac{1}{2}\boldsymbol{g}_j'\boldsymbol{Y}_i \tag{4.6}$$

where

$$\boldsymbol{g}_1 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix} \quad \boldsymbol{g}_2 = \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \end{pmatrix} \quad \boldsymbol{g}_3 = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}. \tag{4.7}$$

Note that the main effect for F_1 corresponds to $j = 1$, the main effect for F_2 corresponds to $j = 2$, and the interaction between F_1 and F_2 corresponds to $j = 3$. Thus, θ_{i1}^{ITT} measures the causal effect of assigning unit i to level 1 of F_1 averaged over the different assigned levels of F_2. The population-level ITT factorial effects are the average of the unit-level ITT factorial effects,

$$\begin{aligned} \theta_1^{\text{ITT}} &= \frac{1}{N}\sum_{i=1}^{N} \theta_{i1}^{\text{ITT}} \\ \theta_2^{\text{ITT}} &= \frac{1}{N}\sum_{i=1}^{N} \theta_{i2}^{\text{ITT}} \\ \theta_3^{\text{ITT}} &= \frac{1}{N}\sum_{i=1}^{N} \theta_{i3}^{\text{ITT}}. \end{aligned} \tag{4.8}$$

We can further decompose the population-level ITT factorial effects by the principal strata,

$$\begin{aligned}
\theta_1^{\text{ITT}} &= p_c\theta_1^{\text{ITT},(c)} + p_n\theta_1^{\text{ITT},(n)} + p_a\theta_1^{\text{ITT},(a)} \\
\theta_2^{\text{ITT}} &= p_c\theta_2^{\text{ITT},(c)} + p_n\theta_2^{\text{ITT},(n)} + p_a\theta_2^{\text{ITT},(a)} \\
\theta_{12}^{\text{ITT}} &= p_c\theta_3^{\text{ITT},(c)} + p_n\theta_3^{\text{ITT},(n)} + p_a\theta_3^{\text{ITT},(a)},
\end{aligned} \tag{4.9}$$

where

$$\begin{aligned}
\theta_1^{\text{ITT},(t)} &= \frac{1}{N_t}\sum_{i\in\mathcal{C}(t)} \theta_{i1}^{\text{ITT}} \\
\theta_2^{\text{ITT},(t)} &= \frac{1}{N_t}\sum_{i\in\mathcal{C}(t)} \theta_{i2}^{\text{ITT}} \\
\theta_3^{\text{ITT},(t)} &= \frac{1}{N_t}\sum_{i\in\mathcal{C}(t)} \theta_{i3}^{\text{ITT}}.
\end{aligned} \tag{4.10}$$

The objective is to estimate the factorial ITT effects in each principal strata. The factorial ITT effects for compliers are of primary interest because the assigned treatment combination agrees with the received treatment combination. Consequently, $\theta_1^{\text{ITT},(c)}$, often called the complier average causal effect (CACE) for F_1, can be interpreted as measuring the effect of *receiving* level 1 of F_1 averaged over the different levels of F_2. This interpretation can be made more compelling by, for instance, blinding the subjects so that effect of assignment is minimized. If unit i is a complier, in order to say that the difference $Y_i((1,1),(1,1)) - Y_i((0,1),(0,1))$ is due entirely to receiving level 1 of factor F_1, it must be the case that $Y_i((1,1),(1,1)) = Y_i((0,1),(1,1))$. This is more believable if the subjects have been blinded. However, because the potential

outcome $Y_i((0,1),(1,1))$ does not actually exist for compliers, this assumption cannot be supported with data.

Finally, our primary estimands of interest are

$$\begin{aligned}\theta_1^{\text{CACE, FP}} &= \theta_1^{\text{ITT},(c)} \\ \theta_2^{\text{CACE, FP}} &= \theta_2^{\text{ITT},(c)} \\ \theta_3^{\text{CACE, FP}} &= \theta_3^{\text{ITT},(c)}. \end{aligned} \tag{4.11}$$

The additional FP in the notation refers to finite-population. Finite-population refers to the actual N units in the experiment and is different from super-population, which we cover in Section 4.3 when we introduce the probabilistic model.

Returning to the steroid study example, $\theta_1^{\text{CACE, FP}}$ is the effect on the compliers of following the weight lifting program and $\theta_3^{\text{CACE, FP}}$ is the interaction between the weight lifting program and the steroids, the additional muscle size added from weight lifting when the subject receives the steroids. Because we are interested in the effect of following the weight lifting program, not simply being assigned to follow it, $\theta_1^{\text{CACE, FP}}$ and $\theta_3^{\text{CACE, FP}}$ are of greater interest than θ_1^{ITT} and θ_3^{ITT}. However, because all units comply with their assignment to steroids, both $\theta_2^{\text{CACE, FP}}$ and θ_2^{ITT} are of potential interest. While $\theta_2^{\text{CACE, FP}}$ is the effect of steroids on the compliers, θ_2^{ITT} is the effect of steroids on all units.

4.2.4 Observed data

Each unit is randomly assigned to one of the four treatment combinations according to a completely randomized assignment mechanism, so that the number of units assigned to each treatment combination is $N/4$. For each unit $i = 1, \ldots, N$, we observe the observed assigned treatment combination Z_i, the received treatment combination W_i^{obs}, and the observed potential outcome Y_i^{obs}, where $W_i^{\text{obs}} = W_i(Z_i)$ and $Y_i^{\text{obs}} = Y_i(Z_i, W_i^{\text{obs}})$. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$, $\boldsymbol{W}^{\text{obs}} = (W_1^{\text{obs}}, \ldots, W_N^{\text{obs}})$ and $\boldsymbol{Y}^{\text{obs}} = (Y_1^{\text{obs}}, \ldots, Y_N^{\text{obs}})$. Note that we only observe one element of $\boldsymbol{W}_i$ and one element of $\boldsymbol{Y}_i$, the fundamental problem of causal inference. If we could observe all elements of the vectors, we would be able to identify the compliance behavior of all units and directly calculate the estimands of interest.

Because the treatment assignment is completely randomized, the treatment assignment is unconfounded, $\mathrm{p}(\boldsymbol{Z} \mid \boldsymbol{Y}) = \mathrm{p}(\boldsymbol{Z})$. This makes the estimation of the ITT factorial effects straightforward. Following Dasgupta et al. (2012), let

$$\bar{Y}(Z) = \frac{1}{N} \sum_{i=1}^{N} Y_i(Z, W_i(Z)) \tag{4.12}$$

be the average of the potential outcomes for assinged treatment combination Z and let $\bar{\boldsymbol{Y}} = (\bar{Y}(Z))_{Z \in \mathscr{F}}$. Then, we can rewrite θ_j^{ITT} as

$$\theta_j^{\text{ITT}} = \frac{1}{2} \boldsymbol{g}_j' \bar{\boldsymbol{Y}} \tag{4.13}$$

and we can estimate θ_j^{ITT} by letting

$$\bar{Y}^{\text{obs}}(Z) = \frac{1}{N/4} \sum_{i\,:\,Z_i = Z} Y_i^{\text{obs}} \tag{4.14}$$

and $\bar{\boldsymbol{Y}}^{\text{obs}} = (\bar{Y}^{\text{obs}}(Z))_{Z \in \mathscr{F}}$. Then $\bar{\boldsymbol{Y}}^{\text{obs}}$ is an unbiased estimate of $\bar{\boldsymbol{Y}}$, $E(\bar{\boldsymbol{Y}}^{\text{obs}}) = \bar{\boldsymbol{Y}}$ and

$$\widehat{\theta_j^{\text{ITT}}} = \frac{1}{2} \boldsymbol{g}_j' \bar{\boldsymbol{Y}}^{\text{obs}}. \tag{4.15}$$

is an unbiased estimate of θ_j^{ITT}. Dasgupta et al. (2012) derived the standard error of $\widehat{\theta_j^{\text{ITT}}}$.

Because not all units comply with their F_1 assignment, the received treatment combination is confounded, $\text{p}(\boldsymbol{W}^{\text{obs}} \,|\, \boldsymbol{Y}) \neq \text{p}(\boldsymbol{W}^{\text{obs}})$. As a result, to estimate of $\theta_j^{\text{CACE, FP}}$ we must rely on a statistical model.

4.3 Principal stratification framework

In this section, we present the framework for estimating the estimands of interest, $\theta_1^{\text{CACE, FP}}$, $\theta_2^{\text{CACE, FP}}$, and $\theta_3^{\text{CACE, FP}}$. The challenge in estimating $\theta_1^{\text{CACE, FP}}$, $\theta_2^{\text{CACE, FP}}$, and $\theta_3^{\text{CACE, FP}}$ is that we cannot generally identify which units are the compliers from the observed data. In order to determine if unit i is a complier, we need to observe $W_{i1}(Z_1)$ when both $Z_1 = 0$ and $Z_1 = 1$, but of course, we can only assign a unit to one value of Z_1. If we know $W_i(1) = 1$, the unit might be a complier or an always-taker. If we know $W_{i1}(0) = 0$, the unit might be a complier or a never-taker. We show how to adapt the traditional principal stratification Bayesian model from Imbens and Rubin (1997) to estimate the estimands of interest in this setting.

4.3.1 Bayesian model

In the Bayesian approach, we put a probability model on the quantities associated with each unit, Z_i, $\boldsymbol{W}_i$, $\boldsymbol{Y}_i$. For each unit, we observe Z_i, $W_i^{\text{obs}} = W_i(Z_i)$, and $Y_i^{\text{obs}} = Y_i(Z_i, W_i^{\text{obs}})$, which means we observe $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$, $\boldsymbol{W}^{\text{obs}} = (W_1^{\text{obs}}, \ldots, W_N^{\text{obs}})$, and $\boldsymbol{Y}^{\text{obs}} = (Y_1^{\text{obs}}, \ldots, Y_N^{\text{obs}})$.

We let $\boldsymbol{Z}$, $\boldsymbol{W}$, and $\boldsymbol{Y}$ be random variables which follow the probability function $f(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y} \,|\, \pi)$, where

$$\begin{aligned} f(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{Y} \,|\, \pi) &= f(\boldsymbol{W}, \boldsymbol{Y} \,|\, \pi) f(\boldsymbol{Z} \,|\, \boldsymbol{W}, \boldsymbol{Y}, \pi) \\ &= f(\boldsymbol{W}, \boldsymbol{Y} \,|\, \pi) f(\boldsymbol{Z}) \\ &= \prod_{i=1}^{N} f(\boldsymbol{W}_i, \boldsymbol{Y}_i \,|\, \pi) f(\boldsymbol{Z}). \end{aligned} \tag{4.16}$$

The second line is true because we are using a completely randomized assignment mechanism and $\boldsymbol{Z}$ does not depend on $\boldsymbol{W}$, $\boldsymbol{Y}$, or any parameters. Hence, $\boldsymbol{Z}$ is independent of $\boldsymbol{W}$ and $\boldsymbol{Y}$. We also assume that $(\boldsymbol{W}_i, \boldsymbol{Y}_i)$, $i = 1, \ldots, N$, are independent conditional on π and let $f(\pi)$ be the prior for π. As a reference, Imbens and Rubin (1997) justify this form of the density function and prior using deFinetti's theorem and exchangeability.

We focus on deriving the posterior distribution for π, conditional on the observed data, $\boldsymbol{Z}$, $\boldsymbol{W}^{\text{obs}}$, and $\boldsymbol{Y}^{\text{obs}}$, by combining the prior distribution of π, $f(\pi)$, with the observed data likelihood. Once we obtain the posterior for π, we can impute $\boldsymbol{W}^{\text{mis}}$ and $\boldsymbol{Y}^{\text{mis}}$ and derive the posterior distributions of the estimands of interest. Here,

$\boldsymbol{W}^{\text{mis}}$ and $\boldsymbol{Y}^{\text{mis}}$ refer to those elements of the matrices $\boldsymbol{W}$ and $\boldsymbol{Y}$ that were not observed.

We first derive the probability function for the observed data by integrating out $\boldsymbol{W}^{\text{mis}}$ and $\boldsymbol{Y}^{\text{mis}}$. We use the probability function for the observed data to obtain the observed data likelihood and then multiply the observed data likelihood by the prior to obtain the posterior distribution of π conditional on the observed data. We also drop $f(\boldsymbol{Z})$ since it does not depend on π and thus, will not affect the posterior distribution of π. The density function for the observed data, $\boldsymbol{W}^{\text{obs}}$ and $\boldsymbol{Y}^{\text{obs}}$, is

$$\begin{aligned} f(\boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}} \,|\, \pi) &= \int\int f(\boldsymbol{W}, \boldsymbol{Y} \,|\, \pi) d\boldsymbol{W}^{\text{mis}} d\boldsymbol{Y}^{\text{mis}} \\ &= \int\int \prod_{i=1}^{N} f(\boldsymbol{W}_i, \boldsymbol{Y}_i \,|\, \pi) d\boldsymbol{W}^{\text{mis}} d\boldsymbol{Y}^{\text{mis}} \\ &= \prod_{i=1}^{N} \int\int f(\boldsymbol{W}_i, \boldsymbol{Y}_i \,|\, \pi) d\boldsymbol{W}_i^{\text{mis}} d\boldsymbol{Y}_i^{\text{mis}} \\ &= \prod_{i=1}^{N} f(W_i^{\text{obs}}, Y_i^{\text{obs}} \,|\, \pi). \end{aligned} \tag{4.17}$$

To find $f(W_i^{\text{obs}}, Y_i^{\text{obs}} \,|\, \pi)$, we carry out the integral $\int\int f(\boldsymbol{W}_i, \boldsymbol{Y}_i \,|\, \pi) d\boldsymbol{W}_i^{\text{mis}} d\boldsymbol{Y}_i^{\text{mis}}$ by first writing $f(\boldsymbol{W}_i, \boldsymbol{Y}_i \,|\, \pi)$ as $f(\boldsymbol{W}_i \,|\, \pi) f(\boldsymbol{Y}_i \,|\, \boldsymbol{W}_i, \pi)$. We next define $f(\boldsymbol{W}_i \,|\, \pi)$ and $f(\boldsymbol{Y}_i \,|\, \boldsymbol{W}_i, \pi)$.

The probability function for the vector of received treatment combinations is defined using the function $\delta(t, \boldsymbol{W}_i)$. The fuction $\delta(t, \boldsymbol{W}_i)$ is 1 if $\boldsymbol{W}_i$ is consistent with principal strata t and 0 otherwise. For instance, if $\boldsymbol{W}_i = c((0,0),(0,1),(1,0),(1,1))$, then unit i is a complier and $\delta(c, \boldsymbol{W}_i) = 1$. We then let

$$f(\boldsymbol{W}_i \,|\, \pi) = \prod_{t \in \{c,n,a\}} \omega_t^{\delta(t,\boldsymbol{W}_i)}, \tag{4.18}$$

where $\omega_t = \Pr(C_i = t)$ is the probability a unit belong to principal strata t, so $\omega_c + \omega_n + \omega_a = 1$.

The probability function for the vector of potential outcomes conditional on the vector of received treatment combinations, $f(\boldsymbol{Y}_i \,|\, \boldsymbol{W}_i, \pi)$, is defined similarly. Let $\eta_{t,z}$ be the parameter that controls the marginal distribution of $Y_i(z, W_i(z))$ given that unit i belongs to principal compliance strata t. Also, let $\eta_{t,\text{assoc}}$ be the parameter that controls the dependence between unit-level potential outcomes for units belonging to principal compliance strata t. Then, let

$$f(\boldsymbol{Y}_i \,|\, \boldsymbol{W}_i, \pi) = \prod_{t \in \{c,n,a\}} \Big(h_t(\boldsymbol{Y}_i \,|\, \eta_{t,\text{assoc}}) \prod_{z \in \mathscr{F}} b_{t,z}(Y_i(z, W_i(z)) \,|\, \eta_{t,z}) \Big)^{\delta(t,\boldsymbol{W}_i)}, \tag{4.19}$$

where $b_{t,z}(Y_i(z, W_i(z)) \,|\, \eta_{t,z}) = f(Y_i(z, W_i(z)) \,|\, C_i = t, \eta_{t,z})$ is the marginal distribution of the potential outcome $Y_i(z, W_i(z))$ given that unit i belongs to principal compliance strata t. Also, $h_t(\boldsymbol{Y}_i \,|\, \eta_{t,\text{assoc}})$ captures the dependence between the potential outcomes. It is defined such that the product of the marginal distributions and $h_t(\boldsymbol{Y}_i \,|\, \eta_{t,\text{assoc}})$ is the joint distribution of $\boldsymbol{Y}_i \,|\, \boldsymbol{W}_i, \pi$. There are then 18 parameters,

$$\begin{aligned}\pi = (&\omega_c,\, \omega_n,\, \omega_a,\\ &\eta_{c,(0,0)},\, \eta_{c,(0,1)},\, \eta_{c,(1,0)},\, \eta_{c,(1,1)},\, \eta_{c,\text{assoc}},\\ &\eta_{n,(0,0)},\, \eta_{n,(0,1)},\, \eta_{n,(1,0)},\, \eta_{n,(1,1)},\, \eta_{n,\text{assoc}},\\ &\eta_{a,(0,0)},\, \eta_{a,(0,1)},\, \eta_{a,(1,0)},\, \eta_{a,(1,1)},\, \eta_{a,\text{assoc}}).\end{aligned} \tag{4.20}$$

The weak exclusion restriction actually allows us to reduce the number of parameters to 16 because $\eta_{t,(0,0)} = \eta_{t,(1,0)}$ and $\eta_{t,(0,1)} = \eta_{t,(1,1)}$ for $t \in \{n, a\}$.

The observed data likelihood for a single unit is the probability function of the observed data, $f(W_i^{\text{obs}}, Y_i^{\text{obs}} \,|\, \pi)$, evaluated at the observed data. The likelihood depends on the assigned treatment combination, the received treatment combination, and the observed potential outcome. The observed data likelihood is derived by integrating out the missing data, which is equivalent to summing over the probabilities of the principal strata consistent with Z_i and W_i^{obs} and multiplying each probability by the conditional density of the observed potential outcome given the principal strata. For example, say we observe $Z_i = (0,0)$, $W_i^{\text{obs}} = (0,0)$ and $Y_i^{\text{obs}} = y_i^{\text{obs}}$. We know unit i is either a complier or never-taker and the observed data likelihood is then

$$f(W_{\text{obs},i}, Y_{\text{obs},i}|\pi) = \omega_c b^i_{c,(0,0)} + \omega_n b^i_{n,(0,0)}, \tag{4.21}$$

where $b^i_{t,(0,0)} = b_{t,(0,0)}(y_i^{\text{obs}} \,|\, \eta_{t,(0,0)})$. We group the units according to the 8 possible combinations of Z_i and $W_i(Z_i)$, such that if $i \in \mathscr{S}(z,w)$, then $Z_i = z$ and $W_i(Z_i) = w$. The 8 sets are

$$\mathscr{S}((0,0),(0,0)),\ \mathscr{S}((0,0),(1,0)),\ \mathscr{S}((1,0),(1,0)),\ \mathscr{S}((1,0),(0,0))$$
$$\mathscr{S}((0,1),(0,1)),\ \mathscr{S}((0,1),(1,1)),\ \mathscr{S}((1,1),(1,1)),\ \mathscr{S}((1,1),(0,1)). \tag{4.22}$$

Then the observed data likelihood is

$$\begin{aligned} f(\boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}} \mid \pi) &= \prod_{i=1}^{N} f(W_i^{\text{obs}}, Y_i^{\text{obs}} \mid \pi) \\ &= \prod_{i \in \mathscr{S}((0,0),(0,0))} (\omega_c b^i_{c,(0,0)} + \omega_n b^i_{n,(0,0)}) \times \prod_{i \in \mathscr{S}((0,0),(1,0))} \omega_a b^i_{a(0,0)} \\ &\times \prod_{i \in \mathscr{S}(1,0),(1,0))} (\omega_c b^i_{c,(1,0)} + \omega_n b^i_{a,(1,0)}) \times \prod_{i \in \mathscr{S}(1,0),(0,0))} \omega_n b^i_{n,(1,0)} \\ &\times \prod_{i \in \mathscr{S}((0,1),(0,1))} (\omega_c b^i_{c,(0,1)} + \omega_n b^i_{n,(0,1)}) \times \prod_{i \in \mathscr{S}((0,1),(1,1))} \omega_a b^i_{a,(0,1)} \\ &\times \prod_{i \in \mathscr{S}((1,1),(1,1))} (\omega_c b^i_{c,(1,1)} + \omega_a b^i_{a,(1,1)}) \times \prod_{i \in \mathscr{S}((1,1),(0,1))} \omega_n b^i_{n,(1,1)}. \end{aligned} \tag{4.23}$$

Note that $\eta_{t,\text{assoc}}$ is not present in the observed data likelihood. Thus if $\eta_{t,\text{assoc}}$ is a priori independent of the other parameters, the posterior distribution for $\eta_{t,\text{assoc}}$ will be the same as the prior distribution and the $\eta_{t,\text{assoc}}$ will still be independent of the other parameters.

We visualize the principal strata associated with each set in Table 4.1.

Table 4.1: **Visualizing principal strata and assigned and received treatment combination:** The rows refer to the assigned treatment combination and the columns to the received treatment combination. Within each cell, we list the principal strata consistent with the assigned and received treatment combination. For instance, the units in $\mathscr{S}((0,0),(0,0))$ are either compliers (c) or never-takers (n).

		w			
		$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
z	$(0,0)$	c, n		a	
	$(0,1)$		c, n		a
	$(1,0)$	n		c, a	
	$(1,1)$		n		c, a

Table 4.1 also provides some intuition into how the model parameters are estimated. Take the first column of Table 4.1, we know that all units in set $\mathscr{S}((1,0),(0,0))$ are never-takers. We can then use those units to estimate the parameters $\eta_{n,(1,0)}$. Just mentioned earlier, under the weak exclusion restriction, we know that $\eta_{n,(1,0)} = \eta_{n,(0,0)}$. We can then use the estimate of $\eta_{n,(0,0)}$ to predict which units in $\mathscr{S}((0,0),(0,0))$, the mixture of compliers and never-takers, are likely never-takers and consequently, which are most likely compliers. Knowing which units are likely never-takers and compliers, then allows us to estimate ω_c, ω_n, ω_a, and $\eta_{c,(0,0)}$. This process is similarly repeated within the other columns.

Once we have derived the posterior distribution of π,

$$f(\pi \,|\, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}}) \propto f(\boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}} \,|\, \pi)\, f(\pi), \tag{4.24}$$

we can use the fact that the posterior distribution of $\theta_j^{\text{CACE, FP}}$ is

$$
\begin{aligned}
f(\theta_j^{\text{CACE, FP}} \,|\, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}}) = \\
&\int\int f(\theta_j^{\text{CACE, FP}} \,|\, \boldsymbol{W}, \boldsymbol{Y})\, f(\boldsymbol{W}^{\text{mis}}, \boldsymbol{Y}^{\text{mis}} \,|\, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})\, d\boldsymbol{W}^{\text{mis}}\, d\boldsymbol{Y}^{\text{mis}} \\
&= \int\int f(\theta_j^{\text{CACE, FP}} \,|\, \boldsymbol{W}, \boldsymbol{Y}) \int f(\boldsymbol{W}^{\text{mis}}, \boldsymbol{Y}^{\text{mis}} \,|\, \pi, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}}) \\
&f(\pi \,|\, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})\, d\pi\, d\boldsymbol{W}^{\text{mis}}\, d\boldsymbol{Y}^{\text{mis}} \\
&= \int\int\int f(\theta_j^{\text{CACE, FP}} \,|\, \boldsymbol{W}, \boldsymbol{Y})\, f(\boldsymbol{W}^{\text{mis}}, \boldsymbol{Y}^{\text{mis}} \,|\, \pi, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}}) \\
&f(\pi \,|\, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})\, d\pi\, d\boldsymbol{W}^{\text{mis}}\, d\boldsymbol{Y}^{\text{mis}}. \qquad (4.25)
\end{aligned}
$$

We return to this equation shortly and show that it yields a simple sampling scheme.

Up to this point, we have only focused on the finite-population parameters $\theta_j^{\text{CACE, FP}}$, $j = 1, 2, 3$. The super-population parameters might also be of interest and these are defined using the model parameters, $\eta_{t,z}$. For instance, in the case where the potential outcomes follow a normal distribution, $\eta_{t,z} = (\mu_{t,z}, \sigma^2_{t,z})$. Let

$$\boldsymbol{\mu}_t = (\mu_{t,(0,0)}, \mu_{t,(0,1)}, \mu_{t,(1,0)}, \mu_{t,(1,1)}) \qquad (4.26)$$

and then the super-population parameters of interest are

$$\theta_j^{\text{CACE, SP}} = \frac{1}{2}\boldsymbol{g}_j'\boldsymbol{\mu}_c \qquad (4.27)$$

for $j = 1, 2, 3$.

4.3.2 Computation

We next review how to derive the posterior distributions for the finite-population estimands and super-population estimands. We first obtain an approximation to the posterior distribution of π, using either the Gibbs sampler or the EM algorithm. The Gibbs sampler provides posterior draws of π while the EM algorithm provides a large-sample approximation of the posterior distribution by finding the maximum likelihood estimate (MLE) of π. In both the Gibbs and EM cases, once we have the posterior distribution of π, we can compute the posterior distributions of the estimands.

Gibbs sampler and data augmentation

We can draw from the posterior distribution of π, $f(\pi \mid \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})$, using the Gibbs sampler. Those draws are then used to directly derive the posterior distributions of the super-population estimands and to impute the missing potential outcomes, which allows for the derivation of the posterior distributions of the finite-population estimands.

We take a data augmentation approach to sampling from $f(\pi \mid \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})$. We treat $\boldsymbol{C}$, the vector of principal strata, as the unobserved data and sample from the joint distribution of π and $\boldsymbol{C}$, $f(\pi, \boldsymbol{C} \mid \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})$. The complete data for the Gibbs sampler is thus $\boldsymbol{C}$, $\boldsymbol{Z}$, $\boldsymbol{W}^{\text{obs}}$, and $\boldsymbol{Y}^{\text{obs}}$. The Gibbs sampler begins with an initial draw of π, $\pi^{(1)}$. At the ith step, we draw from the posterior of $\boldsymbol{C}$ conditional on $\pi = \pi^{(i)}$, $f(\boldsymbol{C} \mid \pi^{(i)}, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})$. Let $\boldsymbol{C}^{(i)}$ be the draw. Next, we draw from the posterior of π conditional on $\boldsymbol{C} = \boldsymbol{C}^{(i)}$, $f(\pi \mid \boldsymbol{C}^{(i)}, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})$. Let $\pi^{(i+1)}$ be the draw. This is repeated for $i = 1, \ldots, L$.

We first derive $f(\boldsymbol{C} \,|\, \pi, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})$ by noting that

$$f(\boldsymbol{C} \,|\, \pi, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}}) = \prod_{i=1}^{N} f(C_i \,|\, \pi, W_i^{\text{obs}}, Y_i^{\text{obs}}). \tag{4.28}$$

Here, $f(C_i = t \,|\, \pi, W_i^{\text{obs}}, Y_i^{\text{obs}})$ gives the probability that unit i belongs to principal strata t. In some cases, we know which principal strata unit i belongs to. For instance, if $i \in \mathscr{S}((0,0),(1,0))$, then unit i is an always-taker and $f(C_i = a \,|\, \pi, W_i^{\text{obs}}, Y_i^{\text{obs}}) = 1$. As before, each of the 8 patterns has its own form. Rather than writing out the probability for each set, we note that if $i \in \mathscr{S}((0,0),(0,0))$, then

$$f(C_i = c \,|\, \pi, W_i^{\text{obs}}, Y_i^{\text{obs}}) = \frac{\omega_c b^i_{c,(0,0)}}{\omega_c b^i_{c,(0,0)} + \omega_n b^i_{n,(0,0)}}. \tag{4.29}$$

The other patterns follow a similar form and once these probabilities are calculated for each unit, it is straightforward to sample $\boldsymbol{C}$.

The next step is to draw from the conditional posterior distribution of π. We assume that $\boldsymbol{C}$ and $\boldsymbol{W}^{\text{obs}}$ are consistent and use the fact that

$$\begin{aligned} f(\pi \,|\, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}}) &\propto f(\pi) f(\boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}} \,|\, \pi) \\ &\propto f(\pi) f(\boldsymbol{C}, \boldsymbol{Y}^{\text{obs}} \,|\, \pi) \\ &\propto f(\pi) \prod_{i=1}^{N} f(C_i, Y_i^{\text{obs}} \,|\, \pi) \\ &\propto f(\pi) \prod_{z \in \mathcal{F}} \prod_{t \in \{c,n,a\}} \prod_{i \in (\mathcal{C}(t) \cap \mathscr{S}(z,\cdot))} \omega_t \, b^i_{t,z}. \end{aligned} \tag{4.30}$$

If we assume

$$(\omega_c, \omega_n, \omega_a),$$

$$\eta_{c,(0,0)}, \eta_{c,(0,1)}, \eta_{c,(1,0)}, \eta_{c,(1,1)},$$

$$\eta_{n,(0,0)}, \eta_{n,(0,1)}, \eta_{n,(1,0)}, \eta_{n,(1,1)},$$

$$\eta_{a,(0,0)}, \eta_{a,(0,1)}, \eta_{a,(1,0)}, \eta_{a,(1,1)},$$

are apriori independent, then the posterior factors into 13 components. This implies that

$$\begin{aligned} f(\omega_c, \omega_n, \omega_a \,|\, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}}) &\propto f(\omega_c, \omega_n, \omega_a)\, \omega_c^{N_c} \omega_n^{N_n} \omega_a^{N_a} \\ f(\eta_{t,z} \,|\, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}}) &\propto f(\eta_{tz}) \prod_{i \in (\mathcal{C}(t) \cap \mathscr{S}(z,\cdot))} b_{t,z}^i. \end{aligned} \tag{4.31}$$

Conjugate priors do exist for this conditional problem. For instance, if $f(\omega_c, \omega_n, \omega_a)$ is a Dirichlet distribution, then the conditional posterior distibution,

$$f(\omega_c, \omega_n, \omega_a \,|\, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}}),$$

is also a Dirichlet distribution. Similarly if we assume the potential outcomes are normally distributed, $\eta_{t,z}$ is the mean and variance for the normal distribution, and $f(\eta_{t,z})$ is normal-inverse gamma, then the conditional posterior distribution, $f(\eta_{t,z} \,|\, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})$, is normal-inverse gamma.

These draws of π can be directly converted into draws of the super-population

estimands. They can also be used to sample from $f(\theta_j^{\text{CACE, FP}} \,|\, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})$. With the posterior draw of π, we can impute $\boldsymbol{W}^{\text{mis}}$ and $\boldsymbol{Y}^{\text{mis}}$ by drawing from the posterior predictive distribution, $f(\boldsymbol{W}^{\text{mis}}, \boldsymbol{Y}^{\text{mis}} \,|\, \pi, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})$. We then calculate $\theta_j^{\text{CACE, FP}}$ based on the complete matrices $\boldsymbol{W}$ and $\boldsymbol{Y}$. In order to impute $\boldsymbol{Y}^{\text{mis}}$, we do need to consider the correlation between the potential outcomes, controlled by $h_t(\boldsymbol{Y}_i \,|\, \eta_{t,\text{assoc}})$. For simplicity, we assume that the potential outcomes within a unit are independent, $h_t(\boldsymbol{Y}_i \,|\, \eta_{t,\text{assoc}}) = 1$, for all principal strata t. In future work, we intend to relax this assumption.

EM algorithm

We can obtain a large-sample approximation of the posterior distribution of π by finding the MLE using the EM algorithm. We then use the information matrix calculated at the MLE to approximate the covariance matrix and assume π follows a multivariate normal distribution. Because this is a large-sample approximation and we are assuming the data overwhelms the prior, we can ignore the prior distribution, $f(\pi)$. As in the Gibbs sampler, the complete data is $\boldsymbol{C}$, $\boldsymbol{Z}$, $\boldsymbol{W}^{\text{obs}}$, and $\boldsymbol{Y}^{\text{obs}}$, where $\boldsymbol{C}$ is unobserved. In the E-step, we take the expectation of the complete data log-likelihood using the current estimates of the parameters. The log-likelihood, $\ell(\pi)$, is derived from the likelihood, $L(\pi)$, and is

$$\begin{aligned}\ell(\pi) &= \log L(\pi) \\ &= \log\Big(\prod_{z\in\mathscr{F}}\prod_{t\in\{c,n,a\}}\prod_{i\in\mathscr{S}(z,\cdot)}(\omega_t b^i_{t,z})^{\mathbf{1}_{C_i=t}}\Big) \\ &= \sum_{z\in\mathscr{F}}\sum_{t\in\{c,n,a\}}\sum_{i\in\mathscr{S}(z,\cdot)}\mathbf{1}_{C_i=t}\cdot(\log\omega_t + \log b^i_{t,z}). \qquad (4.32)\end{aligned}$$

In the E-step, we simply replace $\mathbf{1}_{C_i=t}$ with $E(\mathbf{1}_{C_i=t}\,|\,\pi, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}})$, where

$$E(\mathbf{1}_{C_i=t}\,|\,\pi, \boldsymbol{Z}, \boldsymbol{W}^{\text{obs}}, \boldsymbol{Y}^{\text{obs}}) = f(C_i = t\,|\,\pi, W_i^{\text{obs}}, Y_i^{\text{obs}}). \qquad (4.33)$$

In the M-step, we find the parameter values that maximize the expected log-likelihood. Because the parameters in the log-likelihood are easily separable, the maximization is straightforward. Once we have our large-sample approximation, we can derive posterior distributions of the finite-population and super-populations estimands as before.

4.4 Simulation

Following Imbens and Rubin (1997), we fit the model to simulated data with continuous normally distributed outcomes. We fit the model using both the Gibbs sampler and EM algorithm and compare the frequency properties of both estimates for the finite-population and super-population estimands.

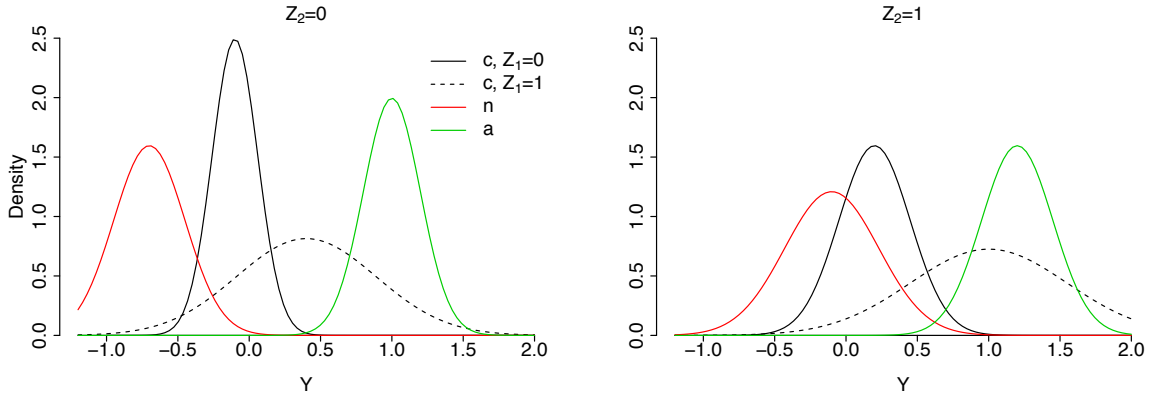
4.4.1 Set-up

We simulate the complete data matrices, $\boldsymbol{W}$ and $\boldsymbol{Y}$, for $N = 400$ units by first independently drawing C_i, which is equivalent to $\boldsymbol{W}_i$, for each unit, where $\omega_c = 0.25$, $\omega_n = 0.45$, and $\omega_a = 0.3$. Given unit i's principal strata, C_i, we draw $\boldsymbol{Y}_i$ from the distributions in Table 4.2.

Table 4.2: **Potential outcomes distributions:** The last four columns give the distributions of the potential outcomes under different principal strata (the rows) and different assigned treatment combinations (the columns).

		$Y \mid C = t$			
t	$P(C_i = t)$	$Z = (0,0)$	$Z = (0,1)$	$Z = (1,0)$	$Z = (1,1)$
c	$\omega_c = 0.25$	$N(-0.1, 0.16)$	$N(0.2, 0.25)$	$N(0.4, 0.49)$	$N(1.0, 0.55)$
n	$\omega_n = 0.45$	$N(-0.7, 0.25)$	$N(-0.1, 0.33)$	$N(-0.7, 0.25)$	$N(-0.1, 0.33)$
a	$\omega_a = 0.3$	$N(1.0, 0.20)$	$N(1.2, 0.25)$	$N(1.0, 0.20)$	$N(1.2, 0.25)$

Note that because of the weak exclusion restriction, the potential outcomes distributions for never-takers and always-takers are the same when $Z = (0, Z_2)$ and $Z = (1, Z_2)$. We visualize the the potenial outcomes distributions in Figure 4.1.



(a) (b)

Figure 4.1: **Potential outcomes distributions:** The distributions are shown for compliers, c, never-takers, n, and always-takers, a, when $Z_2 = 0$ (a) and when $Z_2 = 1$ (b). The distribution for compliers changes with the value of Z_1 (solid black line is $Z_1 = 0$ and dashed black line is $Z_1 = 1$). There is only one potential outcome distribution for never-takers when $Z_2 = 0$ (solid red line) and only one for always-takers when $Z_2 = 0$.

From Table 4.2 and Figure 4.1, we can derive the super-population causal effects, $\theta_j^{\text{CACE, SP}}$, $j = 1, 2, 3$. Using the fact that $\boldsymbol{\mu}_c = (-0.1, 0.2, 0.4, 1.0)$,

$$\begin{aligned}
\theta_1^{\text{CACE, SP}} &= \frac{1}{2}\boldsymbol{g}_1'\boldsymbol{\mu}_c = 0.65 \\
\theta_2^{\text{CACE, SP}} &= \frac{1}{2}\boldsymbol{g}_2'\boldsymbol{\mu}_c = 0.45 \\
\theta_3^{\text{CACE, SP}} &= \frac{1}{2}\boldsymbol{g}_3'\boldsymbol{\mu}_c = 0.15.
\end{aligned} \tag{4.34}$$

F_1 has a positive effect on compliers both when $Z_2 = 0$ and when $Z_2 = 1$ and thus

$\theta_1^{\text{CACE, SP}} > 0$. The effect of F_1 increases when $Z_2 = 1$ and thus $\theta_3^{\text{CACE, SP}} > 0$. In our steroid example, this would imply that weight lifting increases average muscle size when the compliers are not taking steroids and it increases average muscle size even more when the compliers are taking steroids. We also know that F_2 has a positive effect both when $Z_1 = 0$ and when $Z_1 = 1$ and thus $\theta_2^{\text{CACE, SP}} > 0$.

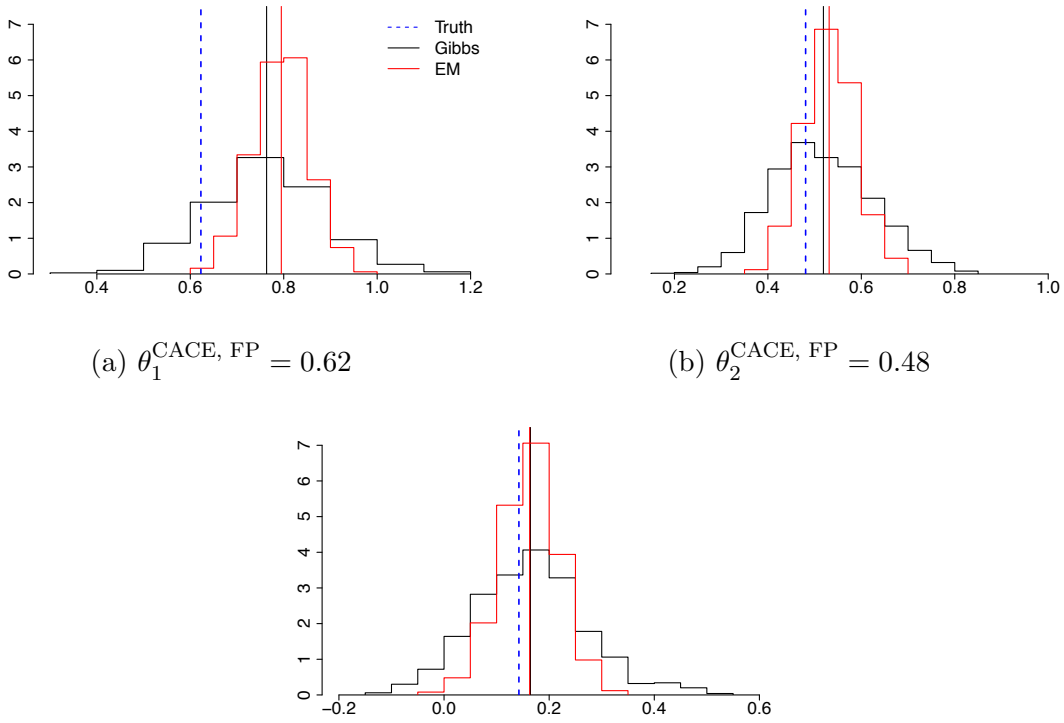
We simulate the observed data by randomly assigning each of the 400 units to one of the four treatment combinations, following a completely randomized assignment mechanism. Using the observed data, we then derive the posterior distributions for the finite-population and super-population estimands using the Gibbs sampler and the EM algorithm. We evaluate the repeated operating characteristics by recording the bias of the posterior mean, the RMSE of the posterior mean, how often the central 95% probability interval covers the true estimand, and the width of the 95% probability interval.

For the prior, we assume $(\omega_c, \omega_n, \omega_a)$ and $\eta_{t,z}$ are independent. We assume a non-informative Dirichlet prior distribution on $(\omega_c, \omega_n, \omega_a)$ such that $f(\omega_c, \omega_n, \omega_a) = \text{Dirichlet}(1, 1, 1)$. We also assume a non-informative prior distribution on $(\mu_{t,z}, \sigma^2_{t,z})$, $f(\mu_{t,z}, \sigma^2_{t,z}) \propto 1/\sigma^2_{t,z}$. Note that this is not a proper prio.

4.4.2 Results

We begin with the results of a single simulated data set. In Figure 4.2, we present histograms for the poterior distributions of the finite populations estimands, derived using both the Gibbs sampler and EM algorithm. For instance, in Figure 4.2(a), the vertical dashed blue line represents the true value of $\theta_1^{\text{CACE, FP}}$ and the black

histogram represents the posterior distribution derived from the Gibbs sampler. The black vertical line is the corresponding posterior mean. The red histogram represent the posterior distribution derived from the large-sample approximation. The red vertical line is the corresponding posterior mean. The posterior distribution derived from the large-sample approximation tends to be much narrowed than the distribution derived from the Gibbs sampler. For example, the 95% interval (0.67, 0.91) from the large sample approximation covers only 66% of the distribution derived from the Gibbs sampler and fails to include the true value of $\theta_1^{\text{CACE, FP}}$. The distributions from the large sample approximation for the other two finite-population estimands, $\theta_2^{\text{CACE, FP}}$ and $\theta_3^{\text{CACE, FP}}$, are also much narrower.



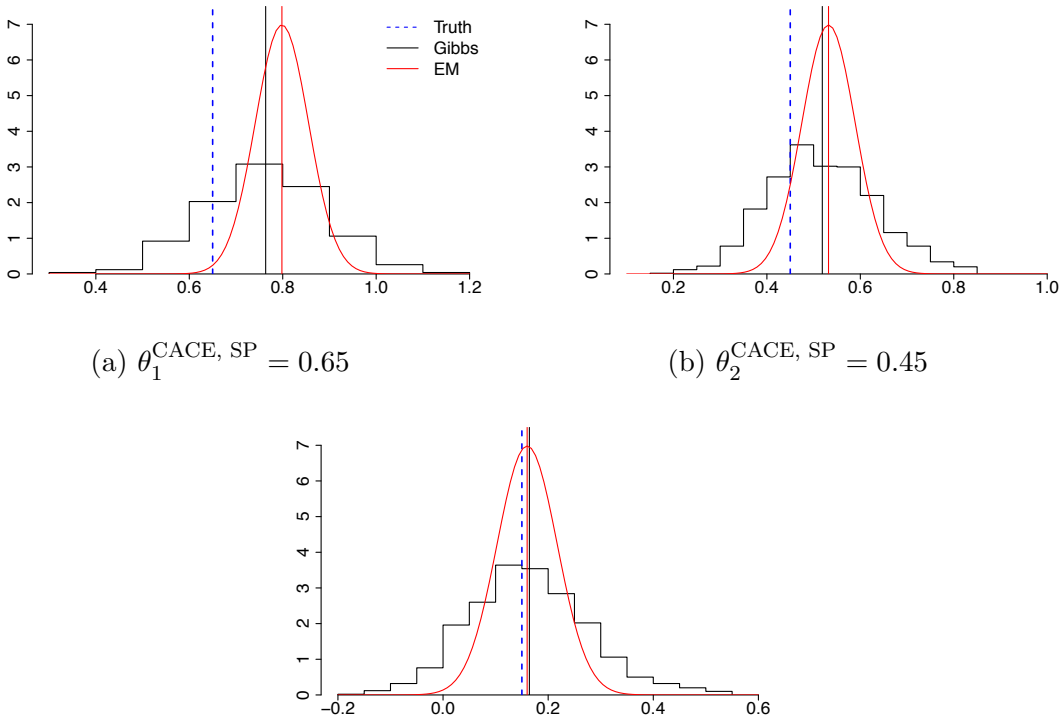
(a) $\theta_1^{\text{CACE, FP}} = 0.62$

(b) $\theta_2^{\text{CACE, FP}} = 0.48$

(c) $\theta_3^{\text{CACE, FP}} = 0.14$

Figure 4.2: **Estimates of finite-population estimands**

In Figure 4.3, we report the posterior distributions of the super-population estimands and the results are similar. The posterior distributions based on the large-sample approximations are much narrower and fail to capture the tail behavior in the distribution derived using the Gibbs sampler.



(a) $\theta_1^{\text{CACE, SP}} = 0.65$

(b) $\theta_2^{\text{CACE, SP}} = 0.45$

(c) $\theta_3^{\text{CACE, SP}} = 0.15$

Figure 4.3: **Estimates of super-population estimands**

We study the repeated operating characteristics of these two methods by repeating this analysis 1000 times. In Tables 4.3 and 4.4, we report the bias of the posterior mean, the RMSE of the posterior mean, how often the central 95% probability interval covers the true estimand, and the width of the 95% probability interval. We see that over many simulations the probability interval from the large-sample approximation fails to cover the true estimand at the nominal 95% level and actually never covers the true value even 80% of the time. We also see that the interval width using the large-sample approximation is much smaller than the interval width using the Gibbs

sampler. Again, this is because the large sample approximation yields a distribution that is too narrow. The repeated-sampling bias and RMSE of the two methods are similar.

Table 4.3: **Simulation results for finite-population estimands**

	Bias	RMSE	Cov. Rate	Mean Width
Gibbs	-0.007	0.117	0.948	0.482
EM	-0.026	0.120	0.721	0.261

(a) $\theta_1^{\text{CACE, FP}}$

	Bias	RMSE	Cov. Rate	Mean Width
Gibbs	-0.009	0.107	0.951	0.439
EM	-0.014	0.110	0.783	0.261

(b) $\theta_2^{\text{CACE, FP}}$

	Bias	RMSE	Cov. Rate	Mean Width
Gibbs	0.021	0.107	0.974	0.922
EM	0.012	0.108	0.770	0.261

(c) $\theta_{12}^{\text{CACE, FP}}$

Table 4.4: **Simulation results for super-population estimands**

	Bias	RMSE	Cov. Rate	Mean Width
Gibbs	-0.006	0.124	0.956	0.509
EM	0.027	0.126	0.727	0.272

(a) $\theta_1^{\text{CACE, SP}} = 0.65$

	Bias	RMSE	Cov. Rate	Mean Width
Gibbs	-0.010	0.115	0.941	0.469
EM	-0.016	0.118	0.766	0.273

(b) $\theta_2^{\text{CACE, SP}} = 0.45$

	Bias	RMSE	Cov. Rate	Mean Width
Gibbs	0.019	0.115	0.962	0.472
EM	0.010	0.115	0.746	0.273

(c) $\theta_{12}^{\text{CACE, SP}} = 0.15$

The simulation confirms that we can successfully estimate the finite-population and super-population estimands and that the Gibbs sampler performs much better than the EM algorithm in terms of approximating the posterior distribution.

4.5 Extensions

We consider two future extensions of the principal stratification framework, allowing non-compliance for both factors and allowing compliance interactions. Both extensions relax one of the earlier assumptions. In this section, we focus on the conceptual issues and leave fitting the models for future work.

4.5.1 Allowing non-compliance for both factors

Up to this point, we have assumed that $W_i(Z_2) = Z_2$, but we could allow units to be either compliers, never-takers, or always-takers with respect to F_2 as well. In the steroid example, this means that units can also fail to comply with their assignment to receive or not to receive the weekly steroid injections. This increases the number of principal strata to $3^2 = 9$. Let

$$C_i \in \{(c,c),(c,n),(c,a),(n,c),(n,n),(n,a),(a,c),(a,n),(a,a)\} \tag{4.35}$$

be the principal strata for the ith unit and, if $C_i = (c,a)$, then unit i is a complier with respect to F_1 and an always-taker with respect to F_2. Also, let $\mathcal{C}(t_1,t_2)$ be the collection of units in principal strata (t_1,t_2). The focus is now to estimate the main effects and interaction for the units in $\mathcal{C}(c,c)$, the units who always comply. Although the number of strata increases when we allow non-compliance for multiple factors, we can still derive the posterior distributions of the estimands. We first modify the exclusion restriction as follows.

4. **Weak exclusion restriction, non-compliance for both factors:** We as-

sume that for units in $\mathcal{C}(n,n)$, $\mathcal{C}(n,a)$, $\mathcal{C}(a,n)$, and $\mathcal{C}(a,a)$, the four potential outcomes are the same. That is,

$$Y_i((0,0),W_i(0,0)) = Y_i((0,1),W_i(0,1)) = Y_i((1,0),W_i(1,0)) = Y_i((1,1),W_i(1,1)). \tag{4.36}$$

Also, for units in $\mathcal{C}(n,c)$ or $\mathcal{C}(a,c)$, the potential outcomes follow the previous exclusion restriction,

$$\begin{aligned} Y_i((0,0),W_i(0,0)) &= Y_i((1,0),W_i(1,0)) \\ Y_i((0,1),W_i(0,1)) &= Y_i((1,1),W_i(1,1)). \end{aligned} \tag{4.37}$$

Finally, for units in $\mathcal{C}(c,n)$ or $\mathcal{C}(c,a)$, the potential outcomes are such that

$$\begin{aligned} Y_i((0,0),W_i(0,0)) &= Y_i((0,1),W_i(0,1)) \\ Y_i((1,0),W_i(1,0)) &= Y_i((1,1),W_i(1,1)). \end{aligned} \tag{4.38}$$

In Table 4.5, we again visualize the principal strata associated with each assigned and received treatment combination.

Table 4.5: **Visualizing principal strata and assigned and received treatment combination when allowing non-compliance for both factors**

		w			
		$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
z	$(0,0)$	(c,c), (c,n), (n,c), (n,n)	(c,a), (n,a)	(a,c), (a,n)	(a,a)
	$(0,1)$	(c,n), (n,n)	(c,c), (c,a), (n,c), (n,a)	(a,n)	(a,c), (a,a)
	$(1,0)$	(n,c), (n,n)	(n,a)	(c,c), (c,n), (a,c), (a,n)	(c,a), (a,a)
	$(1,1)$	(n,n)	(n,c), (n,a)	(c,n), (a,n)	(c,c), (c,a), (a,c), (a,a)

As before, the table provides some intuition for how the model parameters are estimated. Again, focusing on the first column, we know that the units in $\mathscr{S}((1,1),(0,0))$ all belong to $\mathcal{C}(n,n)$ and are never-takers with respect to both F_1 and F_2. Thus, we can estimate the parameter $\eta_{(n,n)}$. Note that according to the modified weak exclusion restriction, all the potential outcomes for units in $\mathcal{C}(n,n)$ follow the same distribution. We can use our estimate of $\eta_{(n,n)}$ to distinguish the (n,n) units in $\mathscr{S}((1,0),(0,0))$ from the (n,c) units and thus estimate $\eta_{(n,c),(1,0)}$. Applying the same logic to the units in $\mathscr{S}((0,1),(0,0))$, we can also estimate $\eta_{(c,n),(0,1)}$. The weak exlcusion restriction also implies that $\eta_{(n,c),(1,0)} = \eta_{(n,c),(0,0)}$ and that $\eta_{(c,n),(0,1)} = \eta_{(c,n),(0,0)}$. Thus, we can apply the estimates of $\eta_{(n,n)}$, $\eta_{(n,c),(0,0)}$, and $\eta_{(c,n),(0,0)}$ to identify the (c,c) units in $\mathscr{S}((0,0),(0,0))$ and estimate $\eta_{(c,c),(0,0)}$. This same logic can be applied to the other three columns, which means that we can estimate the main effects and interaction for the units in $\mathcal{C}(c,c)$. We leave extending this formulation to experiments with more than two factors to future work.

4.5.2 Allowing compliance interactions

Returning to the case with strict compliance for F_2, we allow the assigned level of F_2, Z_{i2}, to affect unit i's compliance behavior regarding F_1. One example of such an interaction is treatment burden, which occurs when subjects become overwhelmed by the requirements of the different treatment assignments and stop complying. As a simple example, suppose in the steroid experiment that a subject complies with his assignment to the weight lifting program if he was not assigned to receive the weekly steroid injections. However, if he was assigned to receive the weekly steroid injections, then he does not follow the weight lifting program because driving to the hospital for the steroid injections takes up too much time. Whether or not such behavior is realistic depends on the context but it can be assessed empirically. For instance, treatment burden implies that the proportion of never-takers is significantly higher for units with $Z_{i2} = 1$ than for units with $Z_{i2} = 0$.

We explore a treatment burden example in a simplified setting. We assume non-compliance is one-sided, so that if a unit is assigned to the inactive level of F_1, there is no way to receive the active level. This means that the regardless of whether $Z_{i2} = 1$ or 0, the unit cannot be an always-taker. We also rule out the possibility that unit i can be a never-taker when $Z_{i2} = 0$ but a complier when $Z_{i2} = 1$. The idea is that treatment burden cannot make a unit more compliant. There are therefore three principal strata, (c, c), (c, n), and (n, n). If $C_i = (c, n)$, equivalently $i \in \mathcal{C}(c, n)$, then unit i is a complier when $Z_{i2} = 0$ and a never-taker when $Z_{i2} = 1$, which is consistent with our treatment burden example. The estimands of interest are the main effects and interaction for those units in $\mathcal{C}(c, c)$.

Even in this simple example, allowing compliance interactions turns out to be more complicated than allowing non-compliance for both factors. In Table 4.6, we again present the principal strata consistent with each assigned and received treatment combination.

Table 4.6: **Visualizing principal strata and assigned and received treatment combination when allowing compliance interactions**

		w			
		$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
z	$(0,0)$	(c,c), (c,n), (n,n)			
	$(0,1)$		(c,c), (c,n), (n,n)		
	$(1,0)$	(n,n)		(c,c), (c,n)	
	$(1,1)$		(c,n), (n,n)		(c,c)

The reason for the complication can be most clearly seen in the third column. The units in $\mathscr{S}((1,0),(1,0))$ are either in $\mathcal{C}(c,c)$ or in $\mathcal{C}(c,n)$ but because no other units receive treatment combination $(1,0)$, it is more difficult to discern which unit is in which strata. The parameters $\eta_{(c,c),(1,0)}$ and $\eta_{(c,n),(1,0)}$ are still technically identified because of the information regarding the proportion of units that belong to each strata. For instance, from the units assigned to treatment combination $(1,0)$, we can estimate the propotion of units in $\mathcal{C}(n,n)$ and, from the units assigned to treatment combination $(1,1)$, we can estimate the proportion of units assigned to $\mathcal{C}(c,c)$. Since there are only three strata, we can then also estimate the proportion of units assigned to $\mathcal{C}(c,n)$. We could then use those proportions to determine which of the two component distributions in the $\mathscr{S}((1,0),(1,0))$ mixture distribution corresponds to the the (c,c) units. This proportion information is clearly part of the likelihood function

and is critical in identifying the estimands. However, in this setting, the posterior distribution of the estimands of interest are multimodal and computation is more difficult.

There are four potential solutions to the computational problem. First, we could collect more data. Second, we could use an informative prior. Up to this point, we have been using improper prior distributions. The computations would remain the same with a proper informative prior but subject matter knowledge would be needed to choose the prior effectively. Third, we could collect covariates that are predictive of compliance behavior. Such covariates could help distinguish the units in $\mathcal{C}(c, c)$ from the units in $\mathcal{C}(c, n)$. Finally, we could consider different Monte Carlo sampling schemes.

4.6 Conclusions

In this chapter, we showed how to apply a Bayesian principal stratification model to a 2^2 factorial experiment in the presence of non-compliance. We believe this is an important application of principal stratification given the prominence of factorial experiments in fields like medicine and education, where the experimental units are often prone to non-compliance. We introduced original notation and assumptions to define the principal strata and identify the model parameters and, in defining the estimands for each principal strata, extended the causal inference framework presented in Dasgupta et al. (2012). Following Imbens and Rubin (1997), we laid out the Bayesian model and associated computational methods and we evaluated the repeated operating characteristics of the posterior means and probability intervals.

We also considered the consequences of allowing subjects to fail to comply with both factor assignments and allowing the compliance behavior of one factor to be influenced by the other factor.

Bibliography

Douglas G Altman. Comparability of randomised groups. *The Statistician*, pages 125–136, 1985.

JD Angrist, Guido W Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91 (434):444–455, 1996.

James O Berger and Robert L Wolpert. The likelihood principle. IMS, 1988.

Shalender Bhasin, Thomas W Storer, Nancy Berman, Carlos Callegari, Brenda Clevenger, Jeffrey Phillips, Thomas J Bunnell, Ray Tricker, Aida Shirazi, and Richard Casaburi. The effects of supraphysiologic doses of testosterone on muscle size and strength in normal men. *New England Journal of Medicine*, 335(1):1–7, 1996.

Allan Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.

James V Bradley. Distribution-free statistical tests. 1968.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, pages 324–345, 1952.

Yuguo Chen, Persi Diaconis, Susan P Holmes, and Jun S Liu. Sequential monte carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, 100(469):109–120, 2005.

Yuguo Chen, Ian H Dinwoodie, and Seth Sullivant. Sequential importance sampling for multiway tables. *The Annals of Statistics*, pages 523–545, 2006.

Jing Cheng and Dylan S. Small. Bounds on causal effects in three-arm trials with noncompliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):815–836, 2006.

David R Cox. Some problems connected with statistical inference. *Ann. Math. Statist*, 29(2):357–372, 1958.

David R Cox. A remark on randomization in clinical trials. *Utilitas Mathematica A*, 21:245–252, 1982.

David R Cox. Discussion of paper by F. Yates. *Journal of the Royal Statistical Society Series A*, 147:451, 1984.

David R Cox and Nancy Reid. *The Theory of the Design of Experiments*. CRC Press, 2000.

Tirthankar Dasgupta, Natesh S. Pillai, and Donald B. Rubin. Causal inference from 2^k factorial designs using the potential outcomes model. 2012.

HA David. The method of paired comparisons. *Charles Griffin, London*, 1988.

Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, pages 363–397, 1998.

Peng Ding. A paradox from randomization-based causal inference. *arXiv preprint arXiv:1402.0142*, 2014.

Bradley Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3): 403–417, 1971.

Ronald A Fisher. Statistical methods and scientific inference. 1956.

Ronald Aylmer Fisher. The arrangement of field experiments. 1926.

Ronald Aylmer Fisher. The design of experiments. 1935.

Constantine E. Frangakis and Donald B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002. ISSN 1541-0420.

MH Gail, WY Tan, and S Piantadosi. Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75(1):57–64, 1988.

Alan Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149, 1992.

Mark E. Glickman. Bayesian locally optimal design of knockout tournaments. *Journal of Statistical Planning and Inference*, 138(7):2117–2127, 2008.

Inge S Helland. Simple counterexamples against the conditionality principle. *The American Statistician*, 49(4):351–356, 1995.

K Hirano, G W Imbens, D B Rubin, and X H Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics (Oxford, England)*, 1 (1):69–88, 2000.

Myles Hollander and Edsel Peña. Nonparametric tests under restricted treatment-assignment rules. *Journal of the American Statistical Association*, 83(404):1144–1151, 1988.

Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *DMKD*. Citeseer, 1997.

FK Hwang. New concepts in seeding knockout tournaments. *American Mathematical Monthly*, pages 235–239, 1982.

Stefano M Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24, 2012.

GW Imbens and DB Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 1997.

Hui Jin and Donald B. Rubin. Principal Stratification for Causal Inference With Extended Partial Compliance. *Journal of the American Statistical Association*, 103(481):101–111, 2008.

John D Kalbfleisch. Sufficiency and conditionality. *Biometrika*, 62(2):251–259, 1975.

Oscar Kempthorne. The design and analysis of experiments. 1952.

Maurice George Kendall. Further contributions to the theory of paired comparisons. *Biometrics*, 11(1):43–62, 1955.

Jack Kiefer. Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, 72(360a):789–808, 1977.

Scott Kirkpatrick, C Daniel Gelatt, Mario P Vecchi, et al. Optimization by simmulated annealing. *Science*, 220(4598):671–680, 1983.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

DV Lindley. *Bayesian Statistics: A Review*. SIAM, 1972.

Roderick J Little, Qi Long, and Xihong Lin. A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics*, 65(2):640–649, 2009.

Q Long, RJA Little, and Xihong Lin. Estimating causal effects in trials involving multi-treatment arms subject to non-compliance: A bayesian framework. *Journal of the Royal Statistical Society: Series C*, 59(3):513–531, 2010.

R Duncan Luce. Individual choice behavior. 1959.

Brendan McCane and Michael Albert. Distance functions for categorical and mixed variables. *Pattern Recognition Letters*, 29(7):986–993, 2008.

Cyrus R Mehta and Nitin R Patel. A network algorithm for performing Fisher's exact test in r× c contingency tables. *Journal of the American Statistical Association*, 78(382):427–434, 1983.

Cyrus R Mehta, Nitin R Patel, and LJ Wei. Constructing exact significance tests with restricted randomization rules. *Biometrika*, 75(2):295–302, 1988.

Luke W Miratrix, Jasjeet S Sekhon, and Bin Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society Series B*, 75(2):369–396, 2013.

Kari Lock Morgan and Donald B Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282, 2012.

William J Morokoff and Russel E Caflisch. Quasi-monte carlo integration. *Journal of computational physics*, 122(2):218–230, 1995.

Frederick Mosteller. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1):3–9, 1951.

Cassandra Wolos Pattanayak. *The Critical Role of Covariate Balance in Causal Inferencewith Randomized Experiments and Observational Studies*. PhD thesis, 2011.

Edwin JG Pitman. Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, pages 322–335, 1938.

Jonathan Raz. Testing for no effect when estimating a smooth function by nonparametric regression: A randomization approach. *Journal of the American Statistical Association*, 85(409):132–138, 1990.

Paul R Rosenbaum. Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574, 1984.

Paul R Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–304, 2002.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Jason Roy, Joseph W Hogan, and Bess H Marcus. Principal stratification with predictors of compliance for randomized trials with 2 active treatments. *Biostatistics (Oxford, England)*, 9(2):277–89, 2008.

Donald B Rubin. Comment. *Journal of the American Statistical Association*, 75(371): 591–593, 1980.

Donald B Rubin. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

Donald B Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26 (1):20–36, 2007.

Allen J Schwenk. What is the correct way to seed a knockout tournament? *American Mathematical Monthly*, pages 140–150, 2000.

Donald T Searls. On the probability of winning with different tournament procedures. *Journal of the American Statistical Association*, 58(304):1064–1081, 1963.

SJ Senn. Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine*, 8(4):467–475, 1989.

RT Smythe. Conditional inference for restricted randomization designs. *The Annals of Statistics*, pages 1155–1161, 1988.

Meir J Stampfer, Julie E Buring, Walter Willett, Bernard Rosner, Kimberly Eberlein, and Charles H Hennekens. The 2× 2 factorial design: Its application to a randomized trial of aspirin and us physicians. *Statistics in Medicine*, 4(2):111–116, 1985.

John F Steiner. Rethinking adherence. *Annals of Internal Medicine*, 157(8):580–585, 2012.

Alisa J Stephens, Eric J Tchetgen Tchetgen, Victor De Gruttola, et al. Flexible covariate-adjusted exact tests of randomized treatment effects with application to a trial of HIV education. *The Annals of Applied Statistics*, 7(4):2106–2137, 2013.

Louis L Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273, 1927.

John W Tukey. Tightening the clinical trial. *Controlled Clinical Trials*, 14(4):266–285, 1993.

L. J. Wei. The adaptive biased coin design for sequential experiments. *The Annals of Statistics*, 6(1):pp. 92–100, 1978.

LJ Wei, RT Smythe, and RL Smith. K-treatment comparisons with restricted randomization rules in clinical trials. *The Annals of Statistics*, pages 265–274, 1986.

LJ Wei, Robert Thomas Smythe, and CR Mehta. Interval estimation with restricted randomization rules. *Biometrika*, 76(2):363–368, 1989.

D. Randall Wilson and Tony R. Martinez. Improved heterogeneous distance functions. *arXiv preprint cs/9701101*, 1997.

MA Yates. The disign and analysis of factorial experiments. 1937.

Min Zhang, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3): 707–715, 2008.

Lu Zheng and Marvin Zelen. Multi-center clinical trials: Randomization and ancillary statistics. *The Annals of Applied Statistics*, pages 582–600, 2008.